Introduction to Nonsmooth Analysis

Theory and Algorithms [1]

Ladislav Lukšan [2] and Jan Vlček [3]

1 Introduction

In this report, basic notions of nonsmooth analysis are investigated together with their application to solution of systems of nonsmooth equations and minimization of nonsmooth functions. All theorems are proved. We use the following classic results without proof exceptionally.

Theorem 1.1 *(Hahn-Banach). Let function $f : R^n \to R$ be positively homogeneous ($f(\lambda x) = \lambda f(x)$ whenever $\lambda \geq 0$) and subadditive ($f(x_1 + x_2) \leq f(x_1) + f(x_2)$). Let $X \subset R^n$ be a subspace and $l : X \to R$ be a linear function such that $l(x) \leq f(x)$, $\forall x \in X$. Then there is a vector $g \in R^n$ such that $g^T x = l(x)$, $\forall x \in X$ and $g^T x \leq f(x)$ $\forall x \in R^n$.*

Theorem 1.2 *(Rademacher). Let function $f : R^n \to R$ be locally Lipschitz (Definition 3.1) in domain $\Omega \in R^n$. Then f is differentiable almost everywhere (set $\{x \in \Omega : \nabla f(x)$ does not exist$\}$ has Lebesgue measure zero).*

Theorem 1.3 *(Karush-Kuhn-Tucker). Let functions $f : R^n \to R$ and $c_i : R^n \to R$, $1 \leq i \leq m$, be continuously differentiable. Let vectors $\{\nabla c_i(x) : c_i(x) = 0, 1 \leq i \leq m\}$ be linearly independent (constraint qualification) at a point $x \in R^n$ which is a local minimum of function f on the set given by constraints $c_i(x) \leq 0$, $i \in I$ and $c_i(x) = 0$, $i \in E$ (where $I \cup E = \{1, \dots, m\}$ and $I \cap E = \emptyset$). Then there is a vector of Lagrange multipliers $\lambda \in R^m$ such that*

$$\nabla f(x) + \sum_{i=1}^{m} \lambda_i \nabla c_i(x) = 0, \tag{1.1}$$

$$\left.\begin{array}{lll} c_i(x) = 0, & & i \in E, \\ c_i(x) \leq 0, & & i \in I, \\ \lambda_i \geq 0, & \lambda_i c_i(x) = 0, & i \in I. \end{array}\right\} \tag{1.2}$$

If the constraint qualification is satisfied, functions f and c_i, $i \in I$, are convex, functions c_i, $i \in E$, are linear and (1.1)-(1.2) hold, then point $x \in R^n$ is the global minimum of function f on the set given by constraints $c_i(x) \leq 0$, $i \in I$ and $c_i(x) = 0$, $i \in E$.

Remark 1.1 If functions $c_i : R^n \to R$, $1 \leq i \leq m$, are linear, constraint qualification is not required.

[1] This work was partly supported by grant GA CR 201/00/0080 and by MURST 1999 Programma di Confinanzimento, Numerical Methods for Evolutionary Problems

[2] Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod vodárenskou věží 2, 182 07 Prague 8, Czech Republic (luksan@uivt.cas.cz)

[3] Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod vodárenskou věží 2, 182 07 Prague 8, Czech Republic (vlcek@uivt.cas.cz)

This report is divided into five parts. Section 1 is an introduction. In Section 2, properties of convex sets and functions are investigated. Section 3 is devoted to Lipschitz continuous functions and mappings. Section 4 concerns systems of semismooth equations. In Section 5, the bundle method for minimizing nonsmooth functions is described. For writing this report, monographs [2] - [8] were used.

2 Convex analysis

2.1 Convex sets

Definition 2.1 *A set $C \in R^n$ is said to be convex, if $x \in C$, $y \in C$ imply*

$$\lambda x + (1-\lambda)y \in C, \tag{2.1}$$

whenever $0 \le \lambda \le 1$.

Remark 2.1 Formula (2.1) can also be written in the form

$$y + \lambda(x-y) \in C.$$

Remark 2.2 It can be easily proved by induction that $x_i \in C$, $\lambda_i \ge 0$, $1 \le i \le m$ and

$$\sum_{i=1}^{m} \lambda_i = 1$$

imply

$$\sum_{i=1}^{m} \lambda_i x_i \in C.$$

Theorem 2.1 *The Intersection of convex sets is a convex set.*

Proof. Let $C = \bigcap_i C_i$, where $C_i \in R^n$ are convex sets. Let $x \in C$, $y \in C$. Then $x \in C_i$ and $y \in C_i$, $\forall i$ and, therefore, $\lambda x + (1-\lambda)y \in C_i$, $\forall i$ whenever $0 \le \lambda \le 1$. It follows that $\lambda x + (1-\lambda)y \in C$. $\diamond$

Theorem 2.2 *A linear combination of convex sets is a convex set.*

Proof. Let $C = \sum_i \lambda_i C_i$, where $C_i \subset R^n$ are convex sets. Let $x \in C$, $y \in C$. Then

$$\lambda x + (1-\lambda)y = \lambda \sum_i \lambda_i x_i + (1-\lambda)\sum_i \lambda_i y_i = \sum_i \lambda_i(\lambda x_i + (1-\lambda)y_i) \triangleq \sum_i \lambda_i z_i$$

holds for $0 \le \lambda \le 1$. Since $x_i \in C_i$, $y_i \in C_i$, we obtain $z_i = \lambda x_i + (1-\lambda)y_i \in C_i$ so that $\lambda x + (1-\lambda)y \in C$. $\diamond$

Definition 2.2 *The convex hull of set $C \subset R^n$ is the set conv C of all convex combinations of points from C, i.e., of all points of the form*

$$y = \sum_{i=1}^{m} \lambda_i x_i, \tag{2.2}$$

where $m \geq 1$, $x_i \in C$, $\lambda_i \geq 0$, $1 \leq i \leq m$, and $\lambda_1 + \ldots + \lambda_m = 1$.

Remark 2.3 *Obviously, $C \in \text{conv } C$.*

Theorem 2.3 *(Caratheodory) Let $y \in \text{conv } C$, where $C \subset R^n$. Then there are at most $n+1$ points $x_i \in C$, $1 \leq i \leq n+1$ such that y is their convex combination.*

Proof. We prove that if (2.2) holds with $m > n+1$, then it is always possible to decrease the number of points in a convex combination. Since m is integer (finite), we get, after a finite number of such steps, a convex combination with at most $n+1$ points. So let

$$y = \sum_{i=1}^{m} \lambda_i x_i,$$

where $m > n+1$, $x_i \in C$, $\lambda_i \geq 0$, $1 \leq i \leq m$, and $\lambda_1 + \ldots + \lambda_m = 1$. Denote

$$\hat{y} = \begin{bmatrix} y \\ 1 \end{bmatrix}, \quad \hat{x}_i = \begin{bmatrix} x_i \\ 1 \end{bmatrix}, \quad 1 \leq i \leq m.$$

Then $\hat{y} \in R^{n+1}$ is a linear combination of vectors $\hat{x}_i \in R^{n+1}$, $1 \leq i \leq m$ (with nonnegative coefficients). Since $m > n+1$, vectors $\hat{x}_i \in R^{n+1}$, $1 \leq i \leq m$, are linearly dependent. Therefore, there are coefficients α_i, $1 \leq i \leq m$, such that at least one of them is nonzero and

$$\sum_{i=1}^{m} \alpha_i \hat{x}_i = 0 \tag{2.3}$$

holds. Since the last elements of vectors $\hat{x}_i$ are units,

$$\sum_{i=1}^{m} \alpha_i = 0$$

holds, so that at least one of these coefficients is negative. From (2.2) and (2.3), we obtain

$$\hat{y} = \sum_{i=1}^{m} \lambda_i \hat{x}_i = \sum_{i=1}^{m} \lambda_i \hat{x}_i + \lambda \sum_{i=1}^{m} \alpha_i \hat{x}_i = \sum_{i=1}^{m} (\lambda_i + \lambda \alpha_i) \hat{x}_i \triangleq \sum_{i=1}^{m} \lambda_i' \hat{x}_i$$

for an arbitrary number $\lambda > 0$. Let

$$\lambda = -\frac{\lambda_j}{\alpha_j} = \min_{\alpha_i < 0} \left(-\frac{\lambda_i}{\alpha_i} \right).$$

Then $\lambda_i' \geq 0$, $1 \leq i \leq m$, $\lambda_j' = 0$ and $\lambda_1' + \ldots + \lambda_m' = 1$ hold so that point y is a convex combination of the $m-1$ points $x_1, \ldots, x_{j-1}, x_{j+1}, \ldots x_m$. ◇

Theorem 2.4 *If the set C is compact, then the set* conv C *is also compact.*

Proof Let $\{x_i\} \subset \text{conv } C$ be a subset such that $x_i \to x \in R^n$. We want to prove that $x \in \text{conv } C$. Since $x_i \in \text{conv } C$, there are, by Theorem 2.3, vectors $y_i^k \in C$ and numbers $\lambda_i^k \geq 0$, $1 \leq k \leq n+1$, $\lambda_i^1 + \ldots + \lambda_i^{n+1} = 1$ such that $x_i = \lambda_i^1 y_i^1 + \ldots + \lambda_i^{n+1} y_i^{n+1}$. Since set C je kompaktn and number $n+1$ is finite, we can choose subsequence $\{\tilde{x}_i\} \subset \{x_i\}$ such that corresponding subsequences $\{\tilde{y}_i^k\} \subset \{y_i^k\}$, $\{\tilde{\lambda}_i^k\} \subset \{\lambda_i^k\}$, $1 \leq k \leq n+1$, are convergent, or $\tilde{y}_i^k \to \tilde{y}^k \in C$, $\tilde{\lambda}_i^k \to \tilde{\lambda}^k \geq 0$, $1 \leq k \leq n+1$, $\tilde{\lambda}^1 + \ldots + \tilde{\lambda}^{n+1} = 1$. Since a chosen subsequence has the same limit as the original convergent sequence, we can write

$$\begin{aligned} x &= \lim_{i\to\infty} x_i = \lim_{i\to\infty} \tilde{x}_i = \lim_{i\to\infty} \left(\sum_{k=1}^{n+1} \tilde{\lambda}_i^k \tilde{y}_i^k \right) \\ &= \sum_{k=1}^{n+1} \left(\lim_{i\to\infty} \tilde{\lambda}_i^k \right) \left(\lim_{i\to\infty} \tilde{y}_i^k \right) = \sum_{k=1}^{n+1} \tilde{\lambda}^k \tilde{y}^k \in \text{conv } C. \end{aligned}$$

□

Definition 2.3 *Let $C \subset R^n$. Then function*

$$d_C(x) = \inf_{y \in C} \|y - x\|$$

is said to be a distance between point x and set C (or the distance function of set C).

Remark 2.4 If set $C \subset R^n$ is compact, then

$$d_C(x) = \min_{y \in C} \|y - x\|.$$

In this report, we focus our attention to compact sets even if most of results hold for more general sets as well.

Theorem 2.5 *Let set $C \subset R^n$ be compact. Then the distance function d_C is Lipschitz continuous with modulus $L = 1$. If C is convex, then d_C is also convex and for each point $x \in R^n$ there is exactly one point $y \in C$ such that*

$$\|y - x\| = d_C(x).$$

Proof. Let $x_1 \in R^n$, $x_2 \in R^n$. By Remark 2.4, there is a point $y \in C$ such that

$$\|y - x_1\| = d_C(x_1).$$

Therefore,

$$d_C(x_2) \leq \|y - x_2\| \leq \|y - x_1\| + \|x_1 - x_2\| = d_C(x_1) + \|x_2 - x_1\|$$

holds, or

$$d_C(x_2) - d_C(x_1) \leq \|x_2 - x_1\|.$$

Since this result is independent of the order of points x_1, x_2, we can write

$$|d_C(x_2) - d_C(x_1)| \le \|x_2 - x_1\|,$$

so that function d_C is Lipschitz continuous with modulus $L = 1$ (Definition 3.1). Let now C be convex, $x_1 \in R^n$, $x_2 \in R^n$. By Remark 2.4, there are points $y_1 \in C$, $y_2 \in C$ such that

$$\begin{aligned} \|y_1 - x_1\| &= d_C(x_1), \\ \|y_2 - x_2\| &= d_C(x_2). \end{aligned}$$

Let us set $y = \lambda_1 y_1 + \lambda_2 y_2$, where $\lambda_1 \ge 0$, $\lambda_2 \ge 0$ and $\lambda_1 + \lambda_2 = 1$. Obviously $y \in C$, so that

$$\begin{aligned} d_C(\lambda_1 x_1 + \lambda_2 x_2) &\le \|y - \lambda_1 x_1 - \lambda_2 x_2\| \le \lambda_1 \|y_1 - x_1\| + \lambda_2 \|y_2 - x_2\| \\ &= \lambda_1 d_C(x_1) + \lambda_2 d_C(x_2). \end{aligned}$$

and d_C is convex in R^n (Definition 2.5). Assume that C is convex and $y_1 \in C$, $y_2 \in C$ are two different points such that $\|y_1 - x\| = d_C(x)$, $\|y_2 - x\| = d_C(x)$. Then

$$\|y_2 - y_1\|^2 = \|(y_2 - x) - (y_1 - x)\|^2 = \|y_2 - x\|^2 + \|y_1 - x\|^2 - 2(y_2 - x)^T(y_1 - x) > 0,$$

so that

$$(y_2 - x)^T(y_1 - x) < d_C^2(x). \tag{2.4}$$

Let us set $y = \frac{1}{2}(y_2 + y_1)$. Since C is convex, we obtain $y \in C$. Furthermore (2.4) implies

$$\|y - x\|^2 = \frac{1}{4}\left(\|y_2 - x\|^2 + \|y_1 - x\|^2 + 2(y_2 - x)^T(y_1 - x)\right) < d_C^2(x),$$

which is a contradiction, since $y \in C$ so that $d_C(x) \le \|y - x\|$ by Remark 2.4. ◇

Theorem 2.6 *Let set C be convex and compact and let $x \notin C$. Then there is a vector $a \in R^n$ and a number $\alpha \in R$ such that*

$$a^T x > \alpha \ge a^T y \quad \forall y \in C.$$

Proof. By Theorem 2.3, there is exactly one vector $\overline{y} \in C$ such that $\|\overline{y} - x\| = d_C(x)$. Let us set $a = x - \overline{y}$ and $\alpha = a^T \overline{y}$. Then

$$a^T x = (x - \overline{y})^T x = (x - \overline{y})^T (x - \overline{y}) + (x - \overline{y})^T \overline{y} = \|x - \overline{y}\|^2 + a^T \overline{y} > \alpha$$

holds, since $x \ne C$ so that $\|x - \overline{y}\| \ne 0$. Inequality $\alpha \ge a^T y$, $\forall y \in C$ will be proven by contradiction. Assume there is a point $y \in C$ such that $\alpha = a^T \overline{y} < a^T y$, and use notation $y(\lambda) = \overline{y} + \lambda b$, where $b = y - \overline{y}$. Obviously, $y(\lambda) \in C$ when $0 \le \lambda \le 1$ (Remark 2.1). Furthermore

$$\|y(\lambda) - x\|^2 = \|\lambda b - a\| = \|a\|^2 - 2\lambda a^T b + \lambda^2 \|b\|^2$$

and

$$\left.\frac{d\|y(\lambda) - x\|^2}{d\lambda}\right|_{\lambda=0} = -2a^T b = -2(a^T y - a^T \overline{y}) < 0$$

hold. Therefore, $\|y(0) - x\|^2 = \|a\|^2$ and there is a number $0 < \overline{\lambda} \le 1$ such that $\|y(\lambda) - x\|^2 < \|a\|^2 = d_C^2(x)$, $\forall 0 < \lambda \le \overline{\lambda}$, which is in contradiction to definition of $d_C(x)$. ◇

Definition 2.4 *Let $C \subset R^n$. Then function*

$$\delta_C(x) = \sup_{y \in C} y^T x$$

is said to be the support function of set C.

Remark 2.5 If $C \subset R^n$ is compact, then

$$\delta_C(x) = \max_{y \in C} y^T x.$$

Theorem 2.7 *Let set $C \subset R^n$ be compact. Then the support function δ_C is positive homogeneous, subadditive and Lipschitz continuous in R_n.*

Proof Using Remark 2.5 for $x \in R^n$ and $\lambda \geq 0$ one has

$$\delta_C(\lambda x) = \max_{y \in C} y^T(\lambda x) = \lambda \max_{y \in C} y^T x = \lambda \delta_C(x),$$

so that function δ_C is positive homogeneous. Similarly, for $x_1 \in R^n$ and $x_2 \in R^n$ one has

$$\delta_C(x_1 + x_2) = \max_{y \in C} y^T(x_1 + x_2) \leq \max_{y \in C} y^T x_1 + \max_{y \in C} y^T x_2 = \delta_C(x_1) + \delta_C(x_2),$$

so that function δ_C is subadditive. The subaddivity implies inequality

$$\delta_C(x_2) \leq \delta_C(x_1) + \delta_C(x_2 - x_1)$$

and since C is compact, there is a constant L such that $\|y\| \leq L \ \forall y \in C$. Therefore, we can write

$$\delta_C(x_2) - \delta_C(x_1) \leq \max_{y \in C} y^T(x_2 - x_1) \leq L\|x_2 - x_1\|.$$

Since this result is independent of the order of points x_1, x_2, one has

$$|\delta_C(x_2) - \delta_C(x_1)| \leq L\|x_2 - x_1\|,$$

so that function δ_C is Lipschitz continuous in R^n (Definition 3.1). □

Theorem 2.8 *Let set $C \subset R^n$ be compact. Then*

$$\delta_C(x) = \delta_{\operatorname{conv} C}(x) \quad \forall x \in R^n.$$

Proof. Since $C \subset \operatorname{conv} C$, then $\delta_C(x) \leq \delta_{\operatorname{conv} C}(x)$, $\forall x \in R^n$ by Remark 2.5. Let $x \in R^n$. By Theorem 2.3, every vector $y \in \operatorname{conv} C$ can be expressed as a convex combination of at most $n+1$ vectors $y_i \in C$, $1 \leq i \leq n+1$. Therefore, we can write

$$\begin{aligned} \delta_{\operatorname{conv} C}(x) &= \max_{y \in \operatorname{conv} C} y^T x = \max \left\{ \sum_{i=1}^{n+1} \lambda_i y_i^T x : y_i \in C, \lambda_i \geq 0, \sum_{i=1}^{n+1} \lambda_i = 1 \right\} \\ &\leq \max_{y \in C} y^T x = \delta_C(x). \end{aligned}$$

□

Theorem 2.9 *Let sets $C_1 \subset R^n$, $C_2 \subset R^n$ are convex and compact. Then $C_1 \subset C_2$ if and only if*

$$\delta_{C_1}(x) \le \delta_{C_2}(x) \quad \forall x \in R^n.$$

Proof. If $C_1 \subset C_2$, then $\delta_{C_1}(x) \le \delta_{C_2}(x)$, $\forall x \in R^n$ by Remark 2.5. Assume $\delta_{C_1}(x) \le \delta_{C_2}(x)$, $\forall x \in R^n$ and that there is a point $\overline{y} \in C_1$ such that $\overline{y} \notin C_2$. Then, by Theorem 2.6, there is a vector $a \in R^n$ and a number $\alpha \in R$ such that

$$a^T \overline{y} > \alpha \ge a^T y \quad \forall y \in C_2.$$

Therefore, one has

$$\delta_{C_1}(a) \ge x^T \overline{y} > \delta_{C_2}(a),$$

which is in contradiction to the assumption. $\diamond$

Theorem 2.10 *Let sets $C_1 \subset R^n$, $C_2 \subset R^n$ be convex and compact. Then*

$$\delta_{C_1+C_2}(x) = \delta_{C_1}(x) + \delta_{C_2}(x).$$

Proof. One has

$$\begin{aligned} \delta_{C_1+C_2}(x) &= \max_{y \in C_1+C_2} y^T x = \max_{\substack{y_1 \in C_1 \\ y_2 \in C_2}} (y_1 + y_2)^T x = \max_{y_1 \in C_1} y_1^T x + \max_{y_2 \in C_2} y_2^T x \\ &= \delta_{C_1}(x) + \delta_{C_2}(x). \end{aligned}$$

$\diamond$

2.2 Convex functions

Definition 2.5 *We say that function $f : R^n \to R$ is convex in a neighborhood of point $x \in R^n$, if there is a number $\varepsilon > 0$ such that f is defined in $B(x, \varepsilon) = \{y : \|y - x\| < \varepsilon\}$ and one has*

$$f(\lambda x_1 + (1 - \lambda) x_2) \le \lambda f(x_1) + (1 - \lambda) f(x_2), \tag{2.5}$$

whenever $x_1 \in B(x, \varepsilon)$, $x_2 \in B(x, \varepsilon)$ and $0 \le \lambda \le 1$. We say that function $f : R^n \to R$ is convex on a convex set $C \subset R^n$, if (2.5) holds whenever $x_1 \in C$, $x_2 \in C$ and $0 \le \lambda \le 1$.

Remark 2.6 Inequality (2.3) can also be written in the form

$$f(x_2 + \lambda(x_1 - x_2)) \le f(x_2) + \lambda(f(x_1) - f(x_2)).$$

Remark 2.7 It can be easily proven by induction that $x_i \in C$, $\lambda_i \ge 0$ and $\sum_{i=1}^m \lambda_i = 1$ imply

$$f\left(\sum_{i=1}^m \lambda_i x_i\right) \le \sum_{i=1}^m \lambda_i f(x_i)$$

if f is convex on C.

Theorem 2.11 *Let function $f : R^n \to R$ be convex in a neighborhood of point $x \in R^n$. Then f is Lipschitz continuous in a neighborhood of x.*

Proof. Since f is convex in a neighborhood of point x, there is a number $\varepsilon > 0$ such that f is defined and convex in $B(x, \varepsilon\sqrt{n+1})$ and, therefore, also in the hypercube

$$\overline{H(x,\varepsilon)} = \{y \in R^n : x_i - \varepsilon \le y_i \le x_i + \varepsilon,\ 1 \le i \le n\} \subset B(x, \varepsilon\sqrt{n+1}).$$

Let $y^{(k)}$, $1 \le k \le 2^n$, be the vertices of this hypercube. Denote

$$M = \max_{1 \le k \le 2^n} f(y^{(k)}).$$

Since every point $\overline{H(x,\varepsilon)}$ can be expressed as a convex combination of vertices $y^{(k)}$, $1 \le k \le 2^n$, the same is also true for points belonging to the neighborhood $B(x,\varepsilon) \subset \overline{H(x,\varepsilon)}$. Let $y \in B(x,\varepsilon)$. Then one has

$$y = \sum_{k=1}^{2^n} \lambda_k y^{(k)}, \quad \sum_{k=1}^{2^n} \lambda_k = 1,$$

where $\lambda_k \ge 0$, $1 \le k \le 2^n$, so that

$$f(y) = f\left(\sum_{k=1}^{2^n} \lambda_k y^{(k)}\right) \le \sum_{k=1}^{2^n} \lambda_k f(y^{(k)}) \le M \sum_{k=1}^{2^n} \lambda_k = M.$$

Therefore, function f is bounded from above on $B(x,\varepsilon)$. Let us choose $y \in B(x,\varepsilon)$ and $y' = 2x - y$. Then $\|y' - x\| = \|x - y\| < \varepsilon$ so that $y' \in B(x,\varepsilon)$. Now convexity implies

$$f(x) = \left(\frac{y+y'}{2}\right) \le \frac{1}{2}(f(y) + f(y'))$$

so that

$$f(y) \ge 2f(x) - f(y') \ge 2f(x) - M$$

and function f is bounded from below on $B(x,\varepsilon)$. Let us set $\delta = \varepsilon/2$ and $m = 2f(x) - M$. Let $z \in B(x,\delta)$, $z' \in B(x,\delta)$ and $z \ne z'$. Denote

$$z'' = z' + \delta\frac{z' - z}{\|z' - z\|} \in B(x,\varepsilon).$$

We obtain

$$z' = \frac{\|z'-z\|}{\delta + \|z'-z\|} z'' + \frac{\delta}{\delta + \|z'-z\|} z$$

by straightforward computation. Now convexity implies

$$\begin{aligned} f(z') - f(z) &\le \frac{\|z'-z\|}{\delta + \|z'-z\|} f(z'') + \frac{\delta}{\delta + \|z'-z\|} f(z) - f(z) \\ &= \frac{\|z'-z\|}{\delta + \|z'-z\|}(f(z'') - f(z)) \le \frac{1}{\delta}\|z'-z\|(M - m). \end{aligned}$$

Since this result is independent of the order of points z and z', we obtain

$$|f(z') - f(z)| \leq \frac{M-m}{\delta}\|z' - z\|$$

so that f is Lipschitz continuous with modulus $L = (M-m)/\delta$ on $B(x, \delta)$. $\diamond$

Lemma 2.1 *Let function $\varphi : R \to R$ be convex on the interval $[a, b]$ and let $a \leq t_1 < t_2 < t_3 \leq b$. Then one has*

$$\frac{\varphi(t_2) - \varphi(t_1)}{t_2 - t_1} \leq \frac{\varphi(t_3) - \varphi(t_1)}{t_3 - t_1} \leq \frac{\varphi(t_3) - \varphi(t_2)}{t_3 - t_2}.$$

Proof. One has

$$t_2 = t_1 + \frac{t_2 - t_1}{t_3 - t_1}(t_3 - t_1),$$

where

$$0 \leq \frac{t_2 - t_1}{t_3 - t_1} \leq 1.$$

Then the convexity of function f (Remark 2.6) implies

$$\varphi(t_2) \leq \varphi(t_1) + \frac{t_2 - t_1}{t_3 - t_1}(\varphi(t_3) - \varphi(t_1)),$$

which proves the left inequality. The right inequality is proved analogously by the expression

$$t_2 = t_3 + \frac{t_2 - t_3}{t_1 - t_3}(t_1 - t_3).$$

Then the convexity of function f implies

$$\varphi(t_2) \leq \varphi(t_3) + \frac{t_2 - t_3}{t_1 - t_3}(\varphi(t_1) - \varphi(t_3)).$$

$\diamond$

Definition 2.6 *We say that function $f : R^n \to R$ has a directional derivative at point $x \in R^n$ and direction $h \in R^n$, if there is the finite limit*

$$f'(x, h) = \lim_{t \downarrow 0} \frac{f(x + th) - f(x)}{t}. \tag{2.6}$$

Theorem 2.12 *Let function $f : R^n \to R$ be convex in a neighborhood of point $x \in R^n$ (so that it is Lipschitz continuous with some modulus L in a neighborhood of this point). Then:*

(a) The directional derivative $f'(x, h)$ exists for every $h \in R^n$ and there is a number ε such that

$$f'(x, h) = \inf_{0 < t \leq \varepsilon / \|h\|} \frac{f(x + th) - f(x)}{t}.$$

(b) *Function $f'(x,\cdot): R^n \to R$ is positively homogeneous, subadditive and Lipschitz continuous with modulus L.*

(c) *Function $f'(\cdot,\cdot): R^n \times R^n \to R$ is upper semicontinuous.*

Proof. (a) Let function f be convex in $B(x,\varepsilon)$. By Lemma 2.1, the function

$$\varphi(t) = \frac{f(x+th)-f(x)}{t}$$

is nondecreasing (left inequality) and bounded from below for $0 < t \le \varepsilon/\|h\|$ (we obtain $-\varphi(-\varepsilon/\|h\|) \le \varphi(t)$ for $0 < t \le \varepsilon/\|h\|$ by coupling both inequalities). Therefore, limit (2.6) exists. The rest of assertion (a) follows from the fact that $\varphi(t)$ is nondecreasing for $0 < t \le \varepsilon/\|h\|$.
(b) Let $\lambda > 0$. Then one has

$$f'(x,\lambda h) = \lim_{t\downarrow 0} \frac{f(x+t\lambda h)-f(x)}{t} = \lambda \lim_{t\downarrow 0} \frac{f(x+t\lambda h)-f(x)}{\lambda t} = \lambda f'(x,h)$$

so that $f'(x,\cdot)$ is positively homogeneous. Furthermore, one has

$$\begin{aligned}
f'(x,h_1+h_2) &= \lim_{t\downarrow 0} \frac{f(x+t(h_1+h_2))-f(x)}{t} \\
&= \lim_{t\downarrow 0} \frac{f\left(\frac{1}{2}(x+2th_1)+\frac{1}{2}(x+2th_2)\right)-f(x)}{t} \\
&\le \lim_{t\downarrow 0} \frac{f(x+2th_1)-f(x)}{2t} + \lim_{t\downarrow 0} \frac{f(x+2th_2)-f(x)}{2t} \\
&= f'(x,h_1)+f'(x,h_2)
\end{aligned}$$

so that $f'(x,\cdot)$ is subadditive. Finally, one has

$$f(x+th_2)-f(x+th_1) \le Lt\|h_2-h_1\|$$

for $t > 0$. Therefore, we can write

$$\lim_{t\downarrow 0} \frac{f(x+th_2)-f(x)}{t} \le \lim_{t\downarrow 0} \frac{f(x+th_1)-f(x)}{t} + L\|h_2-h_1\|$$

so that

$$f'(x,h_2)-f'(x,h_1) \le L\|h_2-h_1\|.$$

By changing the role of vectors h_1 and h_2, we have

$$|f'(x,h_2)-f'(x,h_1) \le L\|h_2-h_1\|,$$

which proves the Lipschitz continuity of $f(x,\cdot)$.
(c) Let $x_i \to x$ and $h_i \to h$. Let us define sequence $t_i \downarrow 0$ so that

$$t_i = \sqrt{\|x_i - x\|} + \frac{1}{i} > 0$$

(we assume that $t_i \le \varepsilon/\|h\|$ without loss of generality). Then one has

$$
\begin{aligned}
f'(x_i,h_i) &= \inf_{0<t\le\varepsilon/\|h\|} \frac{f(x_i+th_i)-f(x_i)}{t} \le \frac{f(x_i+t_ih_i)-f(x_i)}{t_i} \\
&= \frac{f(x+t_ih)-f(x)}{t_i} + \frac{f(x_i+t_ih_i)-f(x+t_ih)}{t_i} + \frac{f(x)-f(x_i)}{t_i}.
\end{aligned}
$$

But

$$
\begin{aligned}
\frac{|f(x_i+t_ih_i)-f(x+t_ih)|}{t_i} &\le \frac{L(\|x_i-x\|+t_i\|h_i-h\|)}{t_i} \\
&\le L(\sqrt{\|x_i-x\|}+\|h_i-h\|) \to 0
\end{aligned}
$$

and

$$
\frac{|f(x_i)-f(x)|}{t_i} \le \frac{L\|x_i-x\|}{t_i} \le L\sqrt{\|x_i-x\|} \to 0.
$$

Therefore, we can write

$$
\limsup_{i\to\infty} f'(x_i,h_i) \le \limsup_{i\to\infty} \frac{f(x+t_ih)-f(x)}{t_i} = \lim_{i\downarrow 0} \frac{f(x+th)-f(x)}{t} = f'(x,h).
$$

$\diamond$

Remark 2.8 Definition 2.6 implies $f'(x,0)=0$ so that, by Theorem 2.12 (b), we obtain

$$
|f'(x,h)| = |f'(x,h)-f'(x,0)| \le L\|h\|.
$$

Definition 2.7 *Let function $f: R^n \to R$ be convex in a neighborhood of point $x \in R^n$. Then the set*

$$
\partial f(x) = \left\{ g \in R^n : f'(x,h) \ge g^T h \ \ \forall h \in R^n \right\}
$$

is the subdifferential of function f at point x. Elements $g \in \partial f(x)$ are subgradients of function f at point x.

Theorem 2.13 *Let function $f: R^n \to R$ be convex in a neighborhood of point $x \in R^n$ (so that it is Lipschitz continuous with some modulus L in a neighborhood of this point). Then:*

(a) The subdifferential $\partial f(x)$ is a nonempty convex compact set such that $\|g\| \le L$, $\forall g \in \partial f(x)$.

(b) For an arbitrary vector $h \in R^n$, one has

$$
f'(x,h) = \max\left\{ g^T h : g \in \partial f(x) \right\}.
$$

(c) If $x_i \to x$, $g_i \in \partial f(x_i)$ and $g_i \to g$, then $g \in \partial f(x)$ (upper semicontinuity).

(d) There exists a number $\varepsilon > 0$ such that for an arbitrary vector $g \in \partial f(x)$, one has

$$f(x+h) - f(x) \ge g^T h \quad \forall h \in B(x, \varepsilon).$$

Proof. (a) By Theorem 2.12 (b), function $f'(x, \cdot)$ is positively homogeneous and subadditive. By the Hahn-Banach theorem, there is a vector $g \in R^n$ such that

$$g^T h \le f'(x, h) \quad \forall h \in R^n$$

so that the subdifferential $\partial f(x)$ is nonempty. Let $g_1 \in \partial f(x)$, $g_2 \in \partial f(x)$ and $\lambda_1 \ge 0$, $\lambda_2 \ge 0$, $\lambda_1 + \lambda_2 = 1$. Then we can write

$$(\lambda_1 g_1 + \lambda_2 g_2)^T h = \lambda_1 g_1^T h + \lambda_2 g_2^T h \le \lambda_1 f'(x, h) + \lambda_2 f'(x, h) = f'(x, h)$$

for any vector $h \in R^n$ so that the subdifferential $\partial f(x)$ is convex. Let $g \in \partial f(x)$. By Definition 2.7 and Remark 2.8, one has

$$\|g\|^2 = g^T g \le f'(x, g) \le L\|g\|$$

or $\|g\| \le L$ so that the subdifferential $\partial f(x)$ is bounded. Let $g_i \in \partial f(x)$ and $g_i \to g$. Then one has

$$g^T h = \lim_{i \to \infty} g_i^T h \le f'(x, h)$$

so that $g \in \partial f(x)$ and the subdifferential $\partial f(x)$ is closed.
(b) By Definition 2.7, one has

$$f'(x, h) \ge \max \left\{ g^T h : g \in \partial f(x). \right\}$$

Assume that

$$f'(x, \overline{h}) > \max \left\{ g^T \overline{h} : g \in \partial f(x) \right\} \tag{2.7}$$

holds for some vector $\overline{h} \in R^n$. Consider the linear function $l(\lambda \overline{h}) \triangleq \lambda f'(x, \overline{h})$ defined on the one-dimensional subspace $\{\lambda \overline{h} : \lambda \in R\} \subset R^n$. Since $f'(x, \cdot)$ is positively homogeneous and subadditive, there is (by Hahn-Banach theorem) a vector $\overline{g} \in R^n$ such that $f'(x, h) \ge \overline{g}^T h$, $\forall h \in R^n$ and $\overline{g}^T(\lambda \overline{h}) = l(\lambda \overline{h}) = \lambda f'(x, \overline{h})$. Therefore, $\overline{g} \in \partial f(x)$ and for $\lambda = 1$ we obtain $f'(x, \overline{h}) = \overline{g}^T \overline{h}$, which is in contradiction to assumption (2.7).
(c) Let $x_i \to x$ and $g_i \to g$, where $g_i \in \partial f(x_i)$. Then, for an arbitrary vector $h \in R^n$, one has

$$g^T h = \lim_{i \to \infty} g_i^T h \le \limsup_{i \to \infty} f'(x_i, h).$$

By Theorem 2.12 (c), function $f'(\cdot, .)$ is upper semicontinuous so that $g^T h \le f'(x, h)$.
(d) Let function f be convex in $B(x, \varepsilon)$ and $g \in \partial f(x)$. By Definition 2.7 and Theorem 2.12 (a), one has

$$g^T h \le f'(x, h) \le \frac{f(x + th) - f(x)}{t}$$

for $0 < t < \overline{t}$ and $h \in B(x, \varepsilon)$. If we choose $t = 1$, we have proved the inequality. $\diamond$

Remark 2.9 Comparing Theorem 2.13 (b) with Remark 2.8, we can see that the directional derivative is the support function of the subdifferential, or

$$f'(x,h) = \delta_{\partial f(x)}(h).$$

Theorem 2.14 *Let function $f : R^n \to R$ be convex in a neighborhood of point $x \in R^n$ and be differentiable at point $x \in R^n$. Then one has*

$$\partial f(x) = \{\nabla f(x)\}.$$

Proof. If f is differentiable at point $x \in R^n$, one has

$$f'(x,h) = (\nabla f(x))^T h.$$

Let $g \in \partial f(x)$. Then, by Definition 2.7, we can write

$$(\nabla f(x))^T h \geq g^T h \quad \forall h \in R^n.$$

For any vector $h \in R^n$, the case $(\nabla f(x))^T h > g^T h$ cannot occur, since it implies $(\nabla f(x))^T(-h) < g^T(-h)$, which is impossible. Therefore, $(\nabla f(x))^T h = g^T h \ \forall h \in R^n$, or $g = \nabla f(x)$. ◇

Theorem 2.15 *Let function $f : R^n \to R$ be convex in a neighborhood of point $x \in R^n$. Then f has a local minimum at point x if and only if*

$$0 \in \partial f(x).$$

Proof. By Theorem 2.12 (a), function $f : R^n \to R$ has a local minimum at point x if and only if $f'(x,h) \geq 0$, $\forall h \in R^n$. Therefore, by Definition 2.7, one has $0 \in \partial f(x)$. If $0 \in \partial f(x)$, then there exists , by Theorem 2.13 (d), a number $\varepsilon > 0$ such that $f(x+h) - f(x) \geq 0$, $\forall h \in B(x,\varepsilon)$ so that f has a local minimum at point x. ◇

Some other properties of subdifferentials will be described in the next section in a more general framework.

3 Lipschitz analysis

3.1 Lipschitz functions

Definition 3.1 *Function $f : R^n \to R$ is said to be Lipschitz continuous in a neighborhood of point $x \in R^n$ (with modulus L), if there is $\varepsilon > 0$ such that one has*

$$|f(x_2) - f(x_1)| \leq L\|x_2 - x_1\|, \tag{3.1}$$

whenever $x_1 \in B(x,\varepsilon)$ and $x_2 \in B(x,\varepsilon)$. Function $f : R^n \to R$ is said to be locally Lipschitz in domain Ω, if it if Lipschitz continuous in a neighborhood of every point $x \in \Omega$.

Definition 3.2 *The generalized (Clarke) directional derivative of function $f : R^n \to R$ at point $x \in R^n$ and direction $h \in R^n$ is defined by the formula*

$$f^0(x,h) = \limsup_{\substack{y \to x \\ t \downarrow 0}} \frac{f(y+th) - f(y)}{t}. \tag{3.2}$$

Remark 3.1 If $f^0(x,h)$ is a generalized directional derivative of function f in the sense of Definition 3.2, there are sequences $x_i \to x$ and $t_i \downarrow 0$ such that

$$f^0(x,h) = \lim_{i \to \infty} \frac{f(x_i + t_i h) - f(x_i)}{t_i}.$$

Theorem 3.1 *Let $f : R^n \to R$ be Lipschitz continuous with modulus L in a neighborhood of point $x \in R^n$. Then:*

(a) Function $f^0(x,\cdot) : R^n \to R$ is positively homogeneous, subadditive and Lipschitz continuous with modulus L.

(b) Function $f^0(\cdot,\cdot) : R^n \times R^n \to R$ is upper semicontinuous.

(c) One has $f^0(x,-h) = (-f)^0(x,h)$, $\forall h \in R^n$.

Proof. (a) Let $\lambda > 0$. Then one has

$$f^0(x,\lambda h) = \limsup_{\substack{y \to x \\ t \downarrow 0}} \frac{f(y+t\lambda h) - f(y)}{t} = \lambda \limsup_{\substack{y \to x \\ t \downarrow 0}} \frac{f(y+t\lambda h) - f(y)}{\lambda t} = \lambda f^0(y,h)$$

so that $f^0(x,\cdot)$ is positively homogeneous. Furthermore, we can write

$$\begin{aligned}
f^0(x,h_1+h_2) &= \limsup_{\substack{y \to x \\ t \downarrow 0}} \frac{f(y+t(h_1+h_2)) - f(y)}{t} \\
&= \limsup_{\substack{y \to x \\ t \downarrow 0}} \left(\frac{f(y+t(h_1+h_2)) - f(y+th_1)}{t} + \frac{f(y+th_1) - f(y)}{t} \right) \\
&\le \limsup_{\substack{y' \to x \\ t \downarrow 0}} \frac{f(y'+th_2) - f(y')}{t} + \limsup_{\substack{y \to x \\ t \downarrow 0}} \frac{f(y+th_1) - f(y)}{t} \\
&= f^0(x,h_2) + f^0(x,h_1),
\end{aligned}$$

where $y' = y + th_1 \to x$ so that $f^0(x,h)$ is subadditive. Since f is Lipschitz continuous with modulus L in a neighborhood of point x, one has

$$f(y+th_2) \le f(y+th_1) + Lt\|h_2 - h_1\|$$

in this neighborhood (see (3.1)) so that

$$
\begin{aligned}
f^0(x,h_2) &= \limsup_{\substack{y\to x\\ t\downarrow 0}} \frac{f(y+th_2)-f(y)}{t} \\
&\le \limsup_{\substack{y\to x\\ t\downarrow 0}} \frac{f(y+th_1)-f(y)}{t} + L\|h_2-h_1\| \\
&= f^0(x,h_1) + L\|h_2-h_1\|
\end{aligned}
$$

or

$$
f^0(x,h_2) - f^0(x,h_1) \le L\|h_1-h_2\|.
$$

Replacing the role of vectors h_1 and h_2, we obtain

$$
|f^0(x,h_2) - f^0(x,h_1)| \le L\|h_2-h_1\|.
$$

Therefore, function $f^0(x,\cdot)$ is Lipschitz continuous with modulus L.
(b) Let $x_i \to x$ and $h_i \to h$. By definition of superior limit, there are sequences $y_i \to x$ and $t_i \downarrow 0$ such that

$$
\begin{aligned}
f^0(x_i,h_i) &\le \frac{f(y_i+t_ih_i)-f(y_i)}{t_i} + \frac{1}{i} \\
&= \frac{f(y_i+t_ih)-f(y_i)}{t_i} + \frac{f(y_i+t_ih_i)-f(y_i+t_ih)}{t_i} + \frac{1}{i}.
\end{aligned}
$$

The Lipschitz continuity implies

$$
\left\| \frac{f(y_i+t_ih_i)-f(y_i+t_ih)}{t_i} \right\| \le L\|h_i-h\|
$$

for sufficiently large indices so that

$$
\limsup_{i\to\infty} f^0(x_i,h_i) \le f^0(x,h) + \lim_{i\to\infty}\left(L\|h_i-h\| + \frac{1}{i}\right) = f^0(x,h).
$$

(c) Obviously

$$
\begin{aligned}
f^0(x,-h) &= \limsup_{\substack{y\to x\\ t\downarrow 0}} \frac{f(y-th)-f(y)}{t} \\
&= \limsup_{\substack{z\to x\\ t\downarrow 0}} \frac{(-f)(z+th)-(-f)(z)}{t} = (-f)^0(x,h)
\end{aligned}
$$

(here $z = y - th$). $\diamond$

Remark 3.2 By Definition 3.2, one has $f^0(x,0)=0$, so that, using Theorem 3.1 (a) we obtain

$$
|f^0(x,h)| = |f^0(x,h) - f^0(x,0)| \le L\|h\|.
$$

Definition 3.3 *Let function $f : R^n \to R$ be Lipschitz continuous in a neighborhood of point $x \in R^n$. Then the set*

$$\partial f(x) = \{g \in R^n : f^0(x,h) \geq g^T h \quad \forall h \in R^n\}$$

is the subdifferential of function f at point x. Elements $g \in \partial f(x)$ are subgradients of function f at point x.

Theorem 3.2 *Let function $f : R^n \to R$ be Lipschitz continuous with modulus L in a neighborhood of point $x \in R^n$. Then:*

(a) The subdifferential $\partial f(x)$ is a nonempty convex compact set such that $\|g\| \leq L$, $\forall g \in \partial f(x)$.

(b) For an arbitrary vector $h \in R^n$, one has

$$f^0(x,h) = \max\left\{g^T h : g \in \partial f(x)\right\}.$$

(c) If $x_i \to x$, $g_i \in \partial f(x_i)$ and $g_i \to g$, then $g \in \partial f(x)$ (upper semicontinuity).

(d) One has $\partial(-f)(x) = -\partial f(x)$.

Proof. (a) By Theorem 3.1 (a), function $f^0(x,\cdot)$ is positively homogeneous and subadditive. By the Hahn-Banach theorem (Theorem 1.1) there is a vector $g \in R^n$ such that

$$g^T h \leq f^0(x,h) \quad \forall h \in R^n$$

so that the subdifferential $\partial f(x)$ is nonempty. Let $g_1 \in \partial f(x)$, $g_2 \in \partial f(x)$ and $\lambda_1 \geq 0$, $\lambda_2 \geq 0$, $\lambda_1 + \lambda_2 = 1$. Then we can write

$$(\lambda_1 g_1 + \lambda_2 g_2)^T h = \lambda_1 g_1^T h + \lambda_2 g_2^T h \leq \lambda_1 f^0(x,h) + \lambda_2 f^0(x,h) = f^0(x,h)$$

for any vector $h \in R^n$ so that the subdifferential $\partial f(x)$ is convex. Let $g \in \partial f(x)$. By Definition 3.3 and Remark 3.2, one has

$$\|g\|^2 = g^T g \leq f^0(x,g) \leq L\|g\|$$

or $\|g\| \leq L$ so that the subdifferential $\partial f(x)$ is bounded. Let $g_i \in \partial f(x)$ and $g_i \to g$. Then one has

$$g^T h = \lim_{i\to\infty} g_i^T h \leq f^0(x,h)$$

so that $g \in \partial f(x)$ and the subdifferential $\partial f(x)$ is closed.
(b) By definition one has

$$f^0(x,h) \geq \max\left\{g^T h : g \in \partial f(x)\right\}.$$

Assume that

$$f^0(x,\overline{h}) > \max\left\{g^T \overline{h} : g \in \partial f(x)\right\} \tag{3.3}$$

holds for some vector $\overline{h} \in R^n$ Consider the linear function $l(\lambda\overline{h}) \triangleq \lambda f^0(x,\overline{h})$ defined on the one-dimensional subspace $\{\lambda\overline{h} : \lambda \in R\} \subset R^n$. Since $f^0(x,\cdot)$ is positively homogeneous and

subadditive, there is (by Hahn-Banach theorem) a vector $\overline{g} \in R^n$ such that $f^0(x,h) \geq \overline{g}^T h$, $\forall h \in R^n$ and $\overline{g}^T(\lambda \overline{h}) = l(\lambda \overline{h}) = \lambda f^0(x,\overline{h})$. Therefore, $\overline{g} \in \partial f(x)$ and for $\lambda = 1$ we obtain $f'(x,\overline{h}) = \overline{g}^T \overline{h}$, which is in contradiction to assumption (3.3).
(c) Let $x_i \to x$ and $g_i \to g$, where $g_i \in \partial f(x_i)$. Then for an arbitrary vector $h \in R^n$, one has

$$g^T h = \lim_{i\to\infty} g_i^T h \leq \limsup_{i\to\infty} f^0(x_i,h).$$

By Theorem 3.1 (b), function $f^0(\cdot,\cdot)$ is upper semicontinuous so that $g^T h \leq f^0(x,h)$.
(d) By Definition 3.3, formula $g \in \partial(-f)(x)$ holds if and only if $(-f)^0(x,h) \geq g^T h$, $\forall h \in R^n$, which is equivalent to $f^0(x,-h) \geq g^T h$, $\forall h \in R^n$, by Theorem 3.1 (c). By Definition 3.3 it means that $-g \in \partial f(x)$. Therefore, $\partial(-f)(x) = -\partial f(x)$. $\diamond$

Remark 3.3 Comparing Theorem 3.2 (b) with Remark 2.5, we can see that the generalized directional derivative is the support function of the subdifferential, or

$$f^0(x,h) = \delta_{\partial f(x)}(h).$$

Theorem 3.3 *Let function $f : R^n \to R$ be continuously differentiable at point $x \in R^n$. Then f is Lipschitz continuous in a neighborhood of point x and one has*

$$\partial f(x) = \{\nabla f(x)\}. \tag{3.4}$$

Proof. If f is continuously differentiable at point x, then gradient $\nabla f(x)$ exists and is bounded in a neighborhood of point x. Therefore, there are numbers $\varepsilon > 0$ and $L > 0$ such that $\|\nabla f(y)\| \leq L$, $\forall y \in B(x,\varepsilon)$. Let $x_1 \in B(x,\varepsilon)$ and $x_2 \in B(x,\varepsilon)$. Then, by the mean value theorem, one has

$$f(x_2) - f(x_1) = (\nabla f(y))^T (x_2 - x_1),$$

where $y \in (x_1,x_2) \subset B(x,\varepsilon)$. Therefore, we can write

$$|f(x_2) - f(x_1)| \leq \|\nabla f(y)\| \, \|x_2 - x_1\| \leq L\|x_2 - x_1\|$$

so that function f is Lipschitz continuous in $B(x,\varepsilon)$. Since function f is continuously differentiable at point x, it follows that $f'(y,h) = (\nabla f(y))^T h$ whenever $y \in B(x,\varepsilon)$. Assume that $x_i \in B(x,\varepsilon)$ and $x_i \to x$. Then, for any $h \in R^n$, one has

$$\begin{aligned} f'(x,h) &= (\nabla f(x))^T h = \lim_{x_i\to x} (\nabla f(x_i))^T h \\ &= \lim_{x_i\to x} f'(x_i,h) = \lim_{\substack{x_i\to x \\ t\downarrow 0}} \frac{f(x_i+th) - f(x_i)}{t} \\ &= \limsup_{\substack{x_i\to x \\ t\downarrow 0}} \frac{f(x_i+th) - f(x_i)}{t} = f^0(x,h) \end{aligned}$$

(if the limit exists, then it is equal to the superior limit). Therefore, we can write $f^0(x,h) = (\nabla f(x))^T h$, $\forall h \in R^n$, so that $\nabla f(x) \in \partial f(x)$. Assume, that $g \in \partial f(x)$ and $g \neq \nabla f(x)$. Then one has $f^0(x,h) = (\nabla f(x))^T h > g^T h$ for some vector $h \in R^n$. But definition of $\partial f(x)$ implies $f^0(x,-h) = -(\nabla f(x))^T h \geq -g^T h$ or (after multiplying by -1) $(\nabla f(x))^T h \leq g^T h$, which is in contradiction to the inequality $(\nabla f(x))^T h > g^T h$. $\diamond$

Remark 3.4 If function $f: R^n \to R$ is Lipschitz continuous in a neighborhood of point $x \in R^n$ and differentiable at this point, then one has

$$\nabla f(x) \in \partial f(x)$$

(since $f^0(x,h) \geq f'(x,h) = (\nabla f(x))^T h, \forall h \in R^n$). Equality (3.4) can be proved only in the case of continuous differentiability.

Theorem 3.4 *Let function $f: R^n \to R$ be convex in a neighborhood of point $x \in R^n$. Then the following statements hold*

(a) $f^0(x,h) = f'(x,h), \forall h \in R^n$.

(b) $\partial f(x) = \left\{ g \in R^n : f'(x,h) \geq g^T h \ \ \forall h \in R^n \right\}$.

Proof. Formula (b) follows immediately form (a) and Definition 3.3. For proving (a), it is sufficient to prove that $f^0(x,h) \leq f'(x,h)$, since the reverse inequality follows immediatelly from Definition 3.2 Let $x_i \to x$ and $t_i \downarrow 0$ so that

$$f^0(x,h) = \lim_{i \to \infty} \frac{f(x_i + t_i h) - f(x_i)}{t_i}$$

(Remark 3.1). Let us set $\bar{t}_i = \max\left(t_i, \sqrt{\|x_i - x\|}\right)$ so that $\|x_i - x\| \leq \bar{t}_i^2$, $t_i \leq \bar{t}_i$ and $\bar{t}_i \to 0$. By Theorem 2.11, function f is Lipschitz continuous (with some modulus L) in a neighborhood of point x (we assume without loss of generality that points x_i, $x_i + \bar{t}_i h$, $x + \bar{t}_i h$ lie in this neighborhood). Using Lemma 2.1 (left inequality) we obtain

$$\begin{aligned}
\frac{f(x_i + t_i h) - f(x_i)}{t_i} &\leq \frac{f(x_i + \bar{t}_i h) - f(x_i)}{\bar{t}_i} \\
&\leq \frac{f(x + \bar{t}_i h) - f(x)}{\bar{t}_i} + \frac{f(x_i + \bar{t}_i h) - f(x + \bar{t}_i h)}{\bar{t}_i} - \frac{f(x_i) - f(x)}{\bar{t}_i} \\
&\leq \frac{f(x + \bar{t}_i h) - f(x)}{\bar{t}_i} + \frac{2L\|x_i - x\|}{\bar{t}_i} \\
&\leq \frac{f(x + \bar{t}_i h) - f(x)}{\bar{t}_i} + 2L\bar{t}_i.
\end{aligned}$$

Passing to limit on both sides of the inequality, we can write

$$f^0(x,h) \leq f'(x,h) + \lim_{\bar{t}_i \to 0} 2L\bar{t}_i = f'(x,h)$$

$\diamond$

Remark 3.5 Theorem 3.4 says that the generalized directional derivative is identical with the usual directional derivative and that the subdifferential by Definition 3.3 coincides with the subdifferential by Definition 2.7 in the convex case.

Equality $f^0(x,h) = f'(x,h)$ is not satisfied in general even if $f'(x,h)$ exists (a nonsmooth concave function is a typical example). However, this equality brings theoretical advantages so that it is reasonable to investigate functions for which it holds.

Definition 3.4 *Function $f : R^n \to R$ is said to be subdifferentially regular at point $x \in R^n$ if, for all $h \in R^n$, the directional derivative $f'(x,h)$ exists and $f^0(x,h) = f'(x,h)$ holds.*

Theorem 3.5 *Continuously differentiable and convex functions are subdifferentially regular. Furthermore, (a) nonnegative linear combinations of subdifferentially regular functions and (b) pointwise maxima of subdifferentially regular functions are subdifferentially regular.*

Proof. A continuously differentiable function is subdifferentially regular by Theorem 3.3 (since one has $f^0(x,h) = \max_{g\in\partial f(x)} g^T h = (\nabla f(x))^T h = f'(x,h)$, $\forall h \in R^n$). A convex function is subdifferentially regular by Theorem 3.4.
(a) It suffices to prove that functions $\lambda_1 f_1$ and $f_1 + f_2$ are subdifferentially regular whenever functions f_1, f_2 are subdifferentially regular and $\lambda_1 \ge 0$ holds. Let $h \in R^n$. If functions f_1, f_2 are subdifferentially regular and $\lambda_1 \ge 0$ holds, then, using Theorem 2.12 (b) and Theorem 3.1 (a), we obtain

$$(\lambda_1 f_1)^0(x,h) = f_1^0(x,\lambda_1 h) = f_1'(x,\lambda_1 h) = (\lambda_1 f_1)'(x,h).$$

Definition 2.6 implies that $(f_1+f_2)'$ exists and $(f_1+f_2)' = f_1' + f_2'$ holds. Definition 3.2 implies $(f_1+f_2)^0 \ge (f_1+f_2)'$. On the other hand, one has

$$\begin{aligned}
(f_1+f_2)^0(x,h) &= \limsup_{\substack{y\to x\\ t\downarrow 0}} \frac{(f_1+f_2)(y+th) - (f_1+f_2)(y)}{t}\\
&= \limsup_{\substack{y\to x\\ t\downarrow 0}} \frac{f_1(y+th) + f_2(y+th) - f_1(y) - f_2(y)}{t}\\
&\le \limsup_{\substack{y\to x\\ t\downarrow 0}} \frac{f_1(y+th) - f_1(y)}{t} + \limsup_{\substack{y\to x\\ t\downarrow 0}} \frac{f_2(y+th) - f_2(y)}{t}\\
&= f_1^0(x,h) + f_2^0(x,h)
\end{aligned}$$

so that

$$(f_1+f_2)' = f_1' + f_2' = f_1^0 + f_2^0 \ge (f_1+f_2)^0,$$

which together with the previous inequality gives $(f_1+f_2)^0 = (f_1+f_2)'$.
(b) It suffices to prove that function $f = \max(f_1, f_2)$ is subdifferentially regular whenever functions f_1, f_2 are subdifferentially regular. If $f_1(x) > f_2(x)$, then $f = f_1$, $f' = f_1'$ and $f^0 = f_1^0 = f_1' = f'$ (the same considerations are used if $f_2(x) > f_1(x)$). Therefore, let $f(x) = f_1(x) = f_2(x)$ and $h \in R^n$. Then

$$\begin{aligned}
f'(x,h) &= \lim_{t\downarrow 0} \frac{f(x+th) - f(x)}{t}\\
&= \lim_{t\downarrow 0} \frac{\max(f_1(x+th), f_2(x+th)) - f(x)}{t}\\
&= \max\left(\lim_{t\downarrow 0} \frac{f_1(x+th) - f_1(x)}{t}, \lim_{t\downarrow 0} \frac{f_2(x+th) - f_2(x)}{t}\right)\\
&= \max(f_1'(x,h), f_2'(x,h))
\end{aligned}$$

so that $f'(x,h)$ exists and $f'(x,h) = \max(f_1'(x,h), f_2'(x,h))$ holds. By Definition 3.2, $f^0(x,h) \geq f'(x,h)$. On the other hand

$$
\begin{aligned}
f^0(x,h) &= \limsup_{\substack{y\to x\\ t\downarrow 0}} \frac{f(y+th)-f(y)}{t} \\
&= \limsup_{\substack{y\to x\\ t\downarrow 0}} \frac{\max(f_1(y+th), f_2(y+th)) - \max(f_1(y), f_2(y))}{t} \\
&\leq \limsup_{\substack{y\to x\\ t\downarrow 0}} \max\left(\frac{f_1(y+th)-f_1(y)}{t}, \frac{f_2(y+th)-f_2(y)}{t}\right) \\
&\leq \max(f_1^0(x,h), f_2^0(x,h)).
\end{aligned}
$$

Therefore, one has

$$f'(x,h) = \max(f_1'(x,h), f_2'(x,h)) = \max(f_1^0(x,h), f_2^0(x,h)) \geq f^0(x,h),$$

which together with the previous inequality gives $f^0(x,h) = f'(x,h)$. $\diamond$

Theorem 3.6 *Let functions $f_1 : R^n \to R$, $f_2 : R^n \to R$ be Lipschitz continuous in a neighborhood of point $x \in R^n$ and $\lambda_1 \in R$. Then:*

(a) $\partial(\lambda_1 f_1)(x) = \lambda_1 \partial f_1(x)$,

(b) $\partial(f_1 + f_2)(x) \subset \partial f_1(x) + \partial f_2(x)$.

If functions f_1, f_2 are subdifferentially regular at point x or one of them is continuously differentiable at point x, then the equality holds in (b).

Proof. (a) If $\lambda_1 \geq 0$, then $(\lambda_1 f_1)^0(x,h) = \lambda_1 f_1^0(x,h)$ so that $\partial(\lambda_1 f_1)(x) = \lambda_1 \partial f_1(x)$ holds by Definition 3.3. In the opposite case, using Theorem 3.2 (d) and the previous result, we get

$$\partial(\lambda_1 f_1)(x) = \partial\left(-|\lambda_1| f_1\right)(x) = -\partial\left(|\lambda_1| f_1\right)(x) = -|\lambda_1| \partial f_1(x) = \lambda_1 \partial f_1(x).$$

(b) Obviously, $(f_1+f_2)^0(x,h) \leq f_1^0(x,h) + f_2^0(x,h)$, $\forall h \in R^n$ (see proof of Theorem 3.5 (a)). Using Remark 3.3 and Theorem 2.10, we obtain

$$\delta_{\partial(f_1+f_2)(x)}(h) \leq \delta_{\partial f_1(x)}(h) + \delta_{\partial f_2(x)}(h) = \delta_{\partial f_1(x)+\partial f_2(x)}(h) \tag{3.5}$$

$\forall h \in R^n$ so that $\partial(f_1+f_2)(x) \subset \partial f_1(x) + \partial f_2(x)$ holds by Theorem 2.9. If functions f_1, f_2 are subdifferentially regular, then $(f_1+f_2)^0 = (f_1+f_2)' = f_1' + f_2' = f_1^0 + f_2^0$ holds by Theorem 3.5 (a) so that the equality holds in (3.5) and in (b) as well. If function f_1 is continuously differentiable at point x, then, using Definition 3.2 and the mean value theorem ($z \in [y, y+th]$), one has

$$
\begin{aligned}
(f_1+f_2)^0(x,h) &= \limsup_{\substack{y\to x\\ t\downarrow 0}} \frac{(f_1+f_2)(y+th) - (f_1+f_2)(y)}{t} \\
&= \lim_{\substack{y\to x\\ t\downarrow 0}} (\nabla f_1(z))^T h + \limsup_{\substack{y\to x\\ t\downarrow 0}} \frac{f_2(y+th)-f_2(y)}{t} \\
&= f_1^0(x,h) + f_2^0(x,h),
\end{aligned}
$$

since $(\nabla f_1(z))^T h \to (\nabla f_1(x))^T h = f_1'(x,h) = f_1^0(x,h)$. $\diamond$

Remark 3.6 *It can be easily proved by induction that*

$$\partial\left(\sum_{i=1}^{m}\lambda_i f_i\right)(x) \subset \sum_{i=1}^{m}\lambda_i \partial f_i(x),$$

where the equality holds if all functions f_i are subdifferentially regular and coefficients λ_i are nonnegative or if all functions f_i but one are continuously differentiable.

Theorem 3.7 *Let function $f : R^n \to R$ be Lipschitz continuous in a neighborhood of point $x \in R^n$, which is its local extremum (minimum or maximum). Then one has*

$$0 \in \partial f(x).$$

Proof. Let $x \in R^n$ be a local minimum of function $f : R^n \to R$. Then necessarily

$$0 \le \limsup_{t\downarrow 0}\frac{f(x+th)-f(x)}{t} \le f^0(x,h)$$

for any $h \in R^n$ so that $0 \in \partial f(x)$ holds by Definition 3.3. If point x is a local maximum of function f, then it is necessarily a local minimum of function $-f$ so that $0 \in \partial(-f)(x)$ and $0 \in \partial f(x)$ holds by Theorem 3.2 (d). $\diamond$

The following mean value theorem is important for further analysis of nonsmooth functions. For simplifying the notation, we use the following definition

$$(\partial f(z))^T v = \left\{g^T v : g \in \partial f(z)\right\}$$

for an arbitrary vector $v \in R^n$.

Theorem 3.8 *Let function $f : R^n \to R$ be locally Lipschitz on an open set Ω containing the line segment $[x,y]$. Then there is a point $z \in (x,y)$ such that*

$$f(y) - f(x) \in (\partial f(z))^T (y-x).$$

Proof. Consider function $\varphi(\lambda) = f(x + \lambda(y-x))$. By assumption, this function is locally Lipschitz on the set containing the interval $[0,1]$. First we prove that

$$\partial\varphi(\lambda) \subset (\partial f(x+\lambda(y-x)))^T (y-x) \tag{3.6}$$

By Theorem 3.2 (a), the sets on both sides of this inclusion are intervals. By Theorem 2.9, it suffices to prove that

$$\delta_{\partial\varphi(\lambda)}(\beta) \le \delta_{(\partial f(x+\lambda(y-x)))^T(y-x)}(\beta) \tag{3.7}$$

for $\beta = 1$ and $\beta = -1$. By Definition 3.2 and Theorem 3.2 (b), one has

$$
\begin{aligned}
\varphi^0(\lambda,\beta) &= \limsup_{\substack{\lambda'\to\lambda \\ t\downarrow 0}} \frac{\varphi(\lambda'+t\beta)-\varphi(\lambda')}{t} \\
&= \limsup_{\substack{\lambda'\to\lambda \\ t\downarrow 0}} \frac{f(x+(\lambda'+t\beta)(y-x))-f(x+\lambda'(y-x))}{t} \\
&\le \limsup_{\substack{y'\to x+\lambda(y-x) \\ t\downarrow 0}} \frac{f(y'+t\beta(y-x))-f(y')}{t} \\
&= f^0(x+\lambda(y-x),\beta(y-x)) \\
&= \max\left\{\beta g^T(y-x) : g\in\partial f(x+\lambda(y-x))\right\}
\end{aligned}
$$

for $\beta=1$ and $\beta=-1$, which gives (3.7) by Remark 2.5 and Remark 3.3 so that also (3.6) holds. Let us set

$$\psi(\lambda)=\varphi(\lambda)-\varphi(0)+\lambda(\varphi(0)-\varphi(1))=f(x+\lambda(y-x))-f(x)+\lambda(f(x)-f(y)).$$

This function is continuous on interval the $[0,1]$ and $\psi(0)=\psi(1)=0$ holds. Therefore, it has to attain a minimum or a maximum in an arbitrary point $\lambda^*\in(0,1)$, which gives $0\in\partial\psi(\lambda^*)$ by Theorem 3.7. If we use Theorem 3.6 and formula (3.6), we obtain

$$0\in\partial\psi(\lambda^*)\subset\partial\varphi(\lambda^*)+\varphi(0)-\varphi(1)\subset(\partial f(x+\lambda^*(y-x)))^T(y-x)+f(x)-f(y),$$

since $\partial(\lambda)=1$, which, after addition of $f(y)-f(x)$ to both sides of this inclusion, gives $f(y)-f(x)\in(\partial f(z))^T(y-x)$ for $z=x+\lambda^*(y-x)\in(x,y)$. $\diamond$

If function $f:R^n\to R$ is locally Lipschitz in a domain Ω, it is differentiable almost everywhere in Ω by the Rademacher theorem so that the set

$$\Omega_f=\{x\in\Omega : \nabla f(x) \text{ does not exist}\}$$

has Lebesgue measure zero. In this case, the subdifferential can also be defined in a different way.

Theorem 3.9 *Let function $f:R^n\to R$ be Lipschitz continuous in a neighborhood of point $x\in R^n$. Then one has*

$$\partial f(x)=\operatorname{conv}\partial_B f(x),$$

where

$$\partial_B f(x)=\left\{\lim_{i\to\infty}\nabla f(x_i) : x_i\to x,\ x_i\notin\Omega_f\right\}.$$

Proof. (a) First we prove that one has

$$f^0(x,h)\le\limsup_{\substack{y\to x \\ y\notin\Omega_f}}\nabla^T f(y)h. \tag{3.8}$$

for an arbitrary $h \in R^n$. Choose $h \in R^n$, $\varepsilon > 0$ arbitrarily and denote by α the right hand side of (3.8). The definition of superior limit implies the existence of a number $\delta > 0$ such that $\nabla^T f(y)h \le \alpha + \varepsilon$ whenever $y \in B(x,\delta)$ and $y \notin \Omega_f$. Without loss of generality, we can assume that f is Lipschitz continuous in $B(x,\delta)$ so that, by the Rademacher theorem, $B(x,\delta) \cap \Omega_f$ has Lebesgue measure zero. Denote

$$L_y = \{y + th : 0 < t < \delta/(2\|h\|)\},$$

so that $L_y \subset B(x,\delta)$ whenever $y \in B(x,\delta/2)$. It follows from the theory of Lebesgue measure that, for almost all points $y \in B(x,\delta/2)$, the set $L_y \cap \Omega_f$ has Lebesgue measure zero. Therefore, for almost all points $y \in B(x,\delta/2)$ and all $t \in (0, \delta/(2\|h\|))$, the following integral exists

$$f(y+th) - f(y) = \int_0^t \nabla^T f(y + \vartheta h) h d\vartheta.$$

Since $\nabla^T f(y+\vartheta h)h \le \alpha + \varepsilon$ whenever $\nabla f(y+\vartheta h)$ exists, we can majorize this integral to obtain

$$f(y+th) - f(y) \le t(\alpha + \varepsilon). \tag{3.9}$$

This inequality holds for almost all points $y \in B(x,\delta/2)$ and all $t \in (0,\delta/(2\|h\|))$. Since function f is continuous, (3.9) holds for all points $y \in B(x,\delta/2)$ and all $t \in (0,\delta/(2\|h\|))$, which gives

$$f^0(x,h) \le \alpha + \varepsilon$$

by Definition 3.2. Since $\varepsilon > 0$ is arbitrary, we obtain (3.8).
(b) Since Ω_f has Lebesgue measure zero, there is at least one sequence $y_i \to x$, $y_i \notin \Omega_f$. By Remark 3.4, $\nabla f(y_i) \in \partial f(y_i)$ holds so that sequence $\{\nabla f(y_i)\}$ is bounded by Theorem 3.2 (a) and there is a convergent subsequence $\{\nabla f(y_i')\} \subset \{\nabla f(y_i)\}$. Therefore, set $\partial_B f(x)$ is nonempty and one has

$$\lim_{i\to\infty} \nabla f(y_i') \in \partial f(x)$$

by Theorem 3.2 (c) so that $\partial_B f(x) \subset \partial f(x)$. Since $\partial f(x)$ is convex, conv $\partial_B f(x) \subset \partial f(x)$ also holds. Since $\partial f(x)$ is compact, sets $\partial_B f(x)$ and conv $\partial_B f(x)$ are also compact. Using Remark 3.3 and inequality (3.8), we obtain

$$\begin{aligned}
\delta_{\partial f(x)}(h) &= f^0(x,h) \le \limsup_{\substack{y\to x \\ y\notin\Omega_f}} \nabla^T f(y) h = \sup_{g\in\partial_B f(x)} g^T h \\
&\le \sup_{g\in \mathrm{conv}\ \partial_B f(x)} g^T h = \delta_{\mathrm{conv}\ \partial_B f(x)}(h)
\end{aligned}$$

for an arbitrary vector $h \in R^n$ so that $\partial f(x) \subset$ conv $\partial_B f(x)$ holds by Theorem 2.9. ◇

3.2 Lipschitz mappings

The approach, used in Theorem 3.9, is a basis for defining the generalized Jacobian of a locally Lipschitz mapping $f : R^n \to R^m$. As in the case of locally Lipschitz function, we define the set

$$\Omega_f = \{x \in \Omega : \mathcal{J} f(x) \text{ does not exist}\},$$

where

$$\mathcal{J}f(x) = \begin{bmatrix} \frac{\partial f_1(x)}{\partial x_1}, & \dots, & \frac{\partial f_1(x)}{\partial x_n} \\ \dots & \dots & \dots \\ \frac{\partial f_m(x)}{\partial x_1}, & \dots, & \frac{\partial f_m(x)}{\partial x_n} \end{bmatrix},$$

which again has Lebesgue measure zero.

Definition 3.5 *Let mapping $f : R^n \to R^m$ be Lipschitz continuous in a neighborhood of point $x \in R^n$. Then the set*

$$\partial f(x) = \text{conv}\ \partial_B f(x),$$

where

$$\partial_B f(x) = \left\{ \lim_{i\to\infty} \mathcal{J}f(x_i) : x_i \to x, x_i \notin \Omega_f \right\},$$

is called the generalized Jacobian of mapping f.

Remark 3.7 Notice that there is a certain inconsistency, since Definition 3.5 differs from Definition 3.3 for $m = 1$ (now we have a row vector). We will, however, stick to this notation since it is common in literature.

Remark 3.8 If mapping $f : R^n \to R^m$ is differentiable at point $x \in R^n$, then Definition 3.5 implies immediately that

$$\mathcal{J}f(x) \in \partial f(x)$$

(it suffices to choose the sequence $x_i = x \to x \notin \Omega_f$).

Theorem 3.10 *Let mapping $f : R^n \to R^m$ be Lipschitz continuous with modulus L in a neighborhood of point $x \in R^n$. Then:*

(a) One has $\partial f(x) \subset [\partial f_1(x), \dots, \partial f_m(x)]^T$, where $\partial f_i(x)$, $1 \le i \le m$, are subdifferentials of functions $f_i : R^n \to R$ (i-th elements of mapping f) at point $x \in R^n$.

(b) The generalized Jacobian $\partial f(x)$ is a nonempty convex compact set such that $\|J\| \le L$, $\forall J \in \partial f(x)$.

(c) If $x_i \to x$, $J_i \in \partial f(x_i)$ and $J_i \to J$, then $J \in \partial f(x)$ (upper semicontinuity).

Proof. (a) follows immediately from Theorem 3.9.
(b) Compactness follows immediately from (a) and Theorem 3.2 (a). Convexity follows directly from Definition 3.5. Nonemptiness follows from the existence of at least one sequence $x_i \to x$, $x_i \notin \Omega_f$, for which $\{\mathcal{J}f(x_i)\}$ converges (the argument is the same as in the proof of Theorem 3.9). Inequality $\|J\| \le L$ follows from Definition 3.5 and from the fact that $\|\mathcal{J}f(x_i)\| \le L$ whenever $\mathcal{J}f(x_i)$ exists.
(c) Assume that $x_i \to x$, $J_i \in \partial f(x_i)$ and $J_i \to J$. By passing to a subsequence, if necessary, we can assume that $x_i \in B(x, 1/(2i))$. If $J \notin \partial f(x)$, then there is a number $\varepsilon > 0$ such that one has

$$J_i \notin \partial f(x) + B(0, \varepsilon).$$

for sufficiently large indices. Since the set $\partial f(x) + B(0,\varepsilon)$ is convex, the inclusion $\partial_B f(x_i) \subset \partial f(x) + B(0,\varepsilon)$ cannot hold (since $J_i \in \text{conv}\partial_B f(x_i) \subset \partial f(x) + B(0,\varepsilon)$ would have to hold in the oposite case). Therefore, there is a matrix $\overline{J}_i \in \partial_B f(x_i)$ such that $\overline{J}_i \notin \partial f(x) + B(0,\varepsilon)$. By Definition 3.5, there is a point $y_i \in B(x_i, 1/(2i)) \subset B(x, 1/i)$ such that $\|\mathcal{J}f(y_i) - \overline{J}_i\| < \varepsilon/2$ so that

$$\mathcal{J}f(y_i) \notin \partial f(x) + B(0,\varepsilon/2). \tag{3.10}$$

By (a), matrices $\overline{J}_i$ and, consequently, matrices $\mathcal{J}f(y_i)$ are uniformly bounded in a neighborhood of point x. Therefore, by passing to a subsequence, if necessary, we can assume that the limit

$$\lim_{i\to\infty} \mathcal{J}f(y_i) = \overline{J}$$

exists. Obviously, $y_i \to x$ (since $y_i \in B(x, 1/i)$), $y_i \notin \Omega_f$ (since $\mathcal{J}f(y_i)$ exists) and $\mathcal{J}f(y_i) \to \overline{J}$. Therefore, $\overline{J} \in \partial f(x)$ holds by Definition 3.5, which is in contradiction to (3.10). ◇

Lemma 3.1 *Let mapping $f : R^n \to R^m$ be Lipschitz continuous in a neighborhood of point $x \in R^n$ and let function $g : R^m \to R$ be continuously differentiable in a neighborhood of point $f(x)$. Then function $\varphi = g \circ f : R^n \to R$ is Lipschitz continuous in a neighborhood of point $x \in R^n$ and one has*

$$\partial\varphi(x) = (\partial f(x))^T \nabla g(f(x)).$$

Proof. Lipschitz continuity of function $g \circ f$ is obvious (it suffices to use Theorem 3.3 and Definition 3.1). Let $J \in \partial_B f(x)$. Then there is a sequence $x_i \to x$, $x_i \notin \Omega_f$ such that $\mathcal{J}f(x_i) \to J$ and, consequently, $\nabla\varphi(x_i) = (\mathcal{J}f(x_i))^T \nabla g(f(x_i)) \to J^T \nabla g(f(x))$. Therefore, $J^T \nabla g(f(x)) \in \partial_B \varphi(x)$ holds, which gives

$$(\partial_B \varphi(x))^T \nabla g(f(x)) \subset \partial_B \varphi(x)$$

On the other hand, let $w \in \partial_B \varphi(x)$. Then there is a sequence $x_i \to x$, $x_i \notin \Omega_\varphi \supset \Omega_f$ such that $\nabla\varphi(x_i) = (\mathcal{J}f(x_i))^T \nabla g(f(x_i)) \to w$. Since Jacobian matrices $\mathcal{J}f(x_i)$ are bounded in a neighborhood of point x by Theorem 3.10 (b), there is a subsequence $\{x_i'\} \subset \{x_i\}$ such that $\mathcal{J}f(x_i') \to J \in \partial_B f(x)$, which together with $(\mathcal{J}f(x_i'))^T \nabla g(f(x_i')) \to w$ gives

$$\partial_B \varphi(x) \subset (\partial_B f(x))^T \nabla g(f(x)).$$

By coupling both inclusions, we obtain $\partial_B \varphi(x) = (\partial_B f(x))^T \nabla g(f(x))$, which gives $\partial\varphi(x) = (\partial f(x))^T \nabla g(f(x))$ by passing to convex hulls. ◇

To formulate the mean value theorem, we use the notation

$$\partial f\left([x,y]\right) = \text{conv} \bigcup_{z\in[x,y]} \partial f(z). \tag{3.11}$$

Lemma 3.2 *Set $\partial f\left([x,y]\right)$ is convex and compact.*

Proof Convexity follows immediately from (3.11). To prove compactness, it suffices, by Theorem 2.4, to show that the set $\bigcup_{z\in[x,y]}\partial f(z)$ is compact. Let $\{J_i\}\subset\bigcup_{z\in[x,y]}\partial f(z)$ be a sequence such that $J_i\to J$. Obviously $J_i\in\partial f(z_i)$, where $z_i\in[x,y]$. Since set $[x,y]$ is compact, there is a subsequence $\{z_i'\}\subset\{z_i\}$ such that $z_i'\to z\in[x,y]$, and a corresponding subsequence $\{J_i'\}\subset\{J_i\}$ such that $J_i'\in\partial f(z_i')$. Since a subsequence has the same limit as the original convergent sequence, we can write $J_i'\to J$, and Theorem 3.10 (c) implies $J\in\partial f(z)\subset\bigcup_{z\in[x,y]}\partial f(z)$. □

Theorem 3.11 *Let mapping $f:R^n\to R^m$ be locally Lipschitz on an open set Ω containing the line segment $[x,y]$. Then one has*

$$f(y)-f(x)\in\partial f\left([x,y]\right)(y-x). \tag{3.12}$$

Proof. By Lemma 3.1, $\partial(v^Tf)(z)=v^T\partial f(z)$ holds for an arbitrary point $z\in(x,y)$ and an arbitrary vector $v\in R^m$. Therefore, we can use Theorem 3.8, which says that, having an arbitrary vector $v\in R^m$, we can found a point $z\in(x,y)$ such that

$$v^T(f(y)-f(x))\in\partial(v^Tf)(z)(y-x)=v^T\partial f(z)(y-x). \tag{3.13}$$

Now we prove (3.12) by contradiction. Assume that $f(y)-f(x)\notin\partial f\left([x,y]\right)(y-x)$. Since the set on the right hand side of (3.12) is convex and compact, there is, by Theorem 2.6, a vector $v\in R^m$ and a number $\alpha\in R$ such that

$$v^T(f(y)-f(x))>\alpha\geq\max_{J\in\partial f([x,y])}v^TJ(y-x).$$

This is in contradiction to (3.13), since (3.13) guarantees that there is an element $J\in\partial f(z)\subset\partial f\left([x,y]\right)$ such that $v^T(f(y)-f(x))=v^TJ(y-x)$. ◇

Theorem 3.12 *Let mapping $f:R^n\to R^m$ be Lipschitz continuous in a neighborhood of point $x\in R^n$ and let function $g:R^m\to R$ be Lipschitz continuous in a neighborhood of point $f(x)$. Then function $\varphi=g\circ f:R^n\to R$ is Lipschitz continuous in a neighborhood of point $x\in R^n$ and one has*

$$\partial\varphi(x)\subset\text{conv}\left(\partial f(x)\right)^T\partial g(f(x))\triangleq\text{conv}\left\{J^Tv:J\in\partial f(x),v\in\partial g(f(x))\right\}, \tag{3.14}$$

where the equality holds in the following particular cases:

(a) Function g is continuously differentiable at point $f(x)$. In this case, we can write

$$\partial\varphi(x)=(\partial f(x))^T\nabla g(f(x)). \tag{3.15}$$

(b) Function g is subdifferentially regular at point $f(x)$ and mapping f is continuously differentiable at point x. In this case, function φ is subdifferentially regular at point x and we can write

$$\partial\varphi(x)=(\mathcal{J}f(x))^T\partial g(f(x)). \tag{3.16}$$

(c) *Function g is subdifferentially regular at point $f(x)$, functions $f_i = e_i^T f$, $1 \le i \le m$, are subdifferentially regular at point x and $v_i \ge 0$, $1 \le i \le m$, holds for an arbitrary element $v \in \partial g(f(x))$. In this case, function φ is subdifferentially regular at point x.*

Proof. Lipschitz continuity of function $g \circ f$ is obvious (double use of Definition 3.1 suffices). Denote by S the set on the right hand side of (3.14). To prove inclusion $\partial\varphi(x) \subset S$, we use Theorem 2.9 and Remark 3.3. Since, by Theorem 2.8, one has

$$\delta_S(h) = \max\left\{v^T Jh : J \in \partial f(x), v \in \partial g(f(x))\right\},$$

for an arbitrary vector $h \in R^n$, it suffices, by Theorem 2.9 and Remark 3.3, to prove that, for an arbitrary vector $h \in R^n$, there is a matrix $J \in \partial f(x)$ and a vector $v \in \partial g(f(x))$ such that

$$\delta_{\partial\varphi(x)}(h) = \varphi^0(x,h) \le v^T Jh. \tag{3.17}$$

By Remark 3.1, we can choose sequences $x_i \to x$ and $t_i \downarrow 0$ such that

$$\varphi^0(x,h) = \lim_{i\to\infty} \frac{\varphi(x_i + t_i h) - \varphi(x_i)}{t_i}.$$

If point $x_i \in R^n$ is sufficiently close to point x and if number $t_i > 0$ is sufficiently small, then points $f(x_i)$ and $f(x_i + t_i h)$ are also sufficiently close to point $f(x)$. Therefore, the assumptions of Theorem 3.8 (applied to function g) are satisfied and, consequently, a point $u_i \in [f(x_i), f(x_i + t_i h)]$ and a subgradient $v_i \in \partial g(u_i)$ exist so that

$$\varphi(x_i + t_i h) - \varphi(x_i) = g(f(x_i + t_i h)) - g(f(x_i)) = v_i^T (f(x_i + t_i h) - f(x_i)).$$

By Theorem 3.11, one has

$$\frac{f(x_i + t_i h) - f(x_i)}{t_i} \in \partial f([x_i, x_i + t_i h])h,$$

which means, by definition (3.11) and Theorem 2.3, that

$$\frac{\varphi(x_i + t_i h) - \varphi(x_i)}{t_i} = v_i^T \frac{f(x_i + t_i h) - f(x_i)}{t_i} = v_i^T \sum_{k=1}^{m+1} \lambda_i^k J_i^k h,$$

where $J_i^k \in \partial f(y_i^k)$, $y_i^k \in [x_i, x_i + t_i h]$, $\lambda_i^k \ge 0$, $k \in [1, m+1]$, $\lambda_i^1 + \ldots + \lambda_i^{m+1} = 1$. For this reason, there is at least one index $k \in [1, m+1]$ such that

$$\frac{\varphi(x_i + t_i h) - \varphi(x_i)}{t_i} \le v_i^T J_i^k h. \tag{3.18}$$

Since $x_i \to x$ and $t_i \downarrow 0$, one has $u_i \to f(x)$ and $y_i^k \to x$. Compactness of subdifferential and generalized Jacobian imply the existence of subsequences $\{x_i'\} \subset \{x_i\}$ and $\{t_i'\} \subset \{t_i\}$ such that the corresponding subsequences $\{v_i'\} \subset \{v_i\}$ and $\{J_i'\} \subset \{J_i^k\}$ converge to v and J. By Theorem 3.2 (c) and Theorem 3.10 (c), one has $v \in \partial g(f(x))$ and $J \in \partial f(x)$ so that (3.18) implies (3.17). Now we will investigate the special cases:
(a) This case is the assertion of Lemma 3.1.
(b) If mapping f is continuously differentiable, then we can represent the set S in the form $S = (\mathcal{J}f(x))^T \partial g(f(x))$ (since set $\partial f(x) = \{\mathcal{J}f(x)\}$ has single element, we need not use the

convex hull). If we use Definition 2.4, Remark 3.3 and regularity of function g (Definition 3.4), we can write

$$\begin{aligned}
\delta_S(h) &= \max_{v\in\partial g(f(x))} v^T \mathcal{J} f(x)h = \max_{v\in\partial g(f(x))} v^T f'(x,h) \\
&= g^0(f(x), f'(x,h)) = g'(f(x), f'(x,h)) \\
&= \lim_{t\downarrow 0} \frac{g(f(x)+tf'(x,h)) - g(f(x))}{t} = \lim_{t\downarrow 0} \left(\frac{g(f(x+th)) - g(f(x))}{t} + T(t) \right),
\end{aligned}$$

where, for sufficiently small t, one has

$$\begin{aligned}
\|T(t)\| &= \frac{\|g(f(x)+tf'(x,h)) - g(f(x+th))\|}{t} \le \frac{L\|f(x)+tf'(x,h) - f(x+th)\|}{t} \\
&= L\left\| f'(x,h) - \frac{f(x+th)-f(x)}{t} \right\|,
\end{aligned}$$

since function g is Lipschitz continuous in some neighborhood of point $f(x)$ (we denoted its modulus by L). Continuous differentiability of mapping f implies that $(f(x+th)-f(x))/t \to f'(x,h)$ so that $T(t) \to 0$ when $t \downarrow 0$. Therefore, we have shown that

$$\varphi'(x,h) = \lim_{t\downarrow 0} \frac{g(f(x+th)) - g(f(x))}{t}$$

exists and $\delta_S(h) = \varphi'(x,h) \le \varphi^0(x,h)$ holds, which by Theorem 2.9 gives $S \subset \partial\varphi(x)$ so that (3.14) implies $\partial\varphi(x) = S$. Moreover, inequality $\varphi^0(x,h) \le \delta_S(h) = \varphi'(x,h) \le \varphi^0(x,h)$ implies regularity of function φ at point x.

(c) Denote

$$S' = \operatorname{conv} \left\{ \sum_{i=1}^m u_i v_i : u_i \in \partial f_i(x), v \in \partial g(f(x)) \right\}.$$

By (3.14) one has $\partial\varphi(x) \subset S$ and by Theorem 3.10 (a) one has $S \subset S'$ so that $\partial\varphi(x) \subset S'$. If functions g and f_i, $1 \le i \le m$, are subdifferentially regular and $v_i \ge 0$, $1 \le i \le m$ holds, we can write

$$\begin{aligned}
\delta_{S'}(h) &= \max \left\{ \sum_{i=1}^m v_i u_i^T h : u_i \in \partial f_i(x), v \in \partial g(f(x)) \right\} \\
&\le \max \left\{ \sum_{i=1}^m v_i \max_{u_i\in\partial f_i(x)} u_i^T h : v \in \partial g(f(x)) \right\} \\
&= \max \left\{ \sum_{i=1}^m v_i f_i^0(x,h) : v \in \partial g(f(x)) \right\} \\
&= \max \left\{ \sum_{i=1}^m v_i f_i'(x,h) : v \in \partial g(f(x)) \right\} \\
&= g^0(f(x), f'(x,h)) = g'(f(x), f'(x,h)).
\end{aligned}$$

The rest of the proof is the same as that of assertion (b). We obtain $\delta_{S'}(h) = \varphi'(x,h) \le \varphi^0(x,h)$, which gives $S' \subset \partial\varphi(x)$ by Theorem 2.9 so that $\partial\varphi(x) \subset S \subset S'$ implies $\partial\varphi(x) = S = S'$.

Moreover, inequality $\varphi^0(x,h) \le \delta_S(h) \le \delta_{S'}(h) = \varphi'(x,h) \le \varphi^0(x,h)$ implies regularity of function φ at point x. $\diamond$

Corollary 3.1 *If the assumptions of Theorem 3.12 are satisfied, then one has*

$$\partial\varphi(x) \subset \operatorname{conv}\left\{\sum_{i=1}^{m} u_i v_i : u_i \in \partial f_i(x), v \in \partial g(f(x))\right\}, \tag{3.19}$$

where the equality holds in the following particular cases:

(a) Function g is continuously differentiable at point $f(x)$ and $m = 1$.

(b) Function g is subdifferentially regular at point $f(x)$, functions f_i, $1 \le i \le m$, are continuously differentiable at point x.

(c) Function g is subdifferentially regular at point $f(x)$, functions f_i, $1 \le i \le m$, are subdifferentially regular at point x and for an arbitrary element $v \in \partial g(f(x))$ one has $v_i \ge 0$, $1 \le i \le m$.

Proof. It suffices to use Theorem 3.12 and some considerations (e.g. $S \subset S'$) from its proof. $\diamond$

Corollary 3.2 *Let functions $f_1 : R^n \to R$, $f_2 : R^n \to R$ be Lipschitz continuous in a neighborhood of point $x \in R^n$. Then function $\varphi = f_1 f_2$ is Lipschitz continuous in a neighborhood of point x. Moreover, denoting*

$$f = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix}, \quad P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

one has

$$\partial\varphi(x) = (\partial f(x))^T P f(x) \subset \partial f_1(x) f_2(x) + f_1(x)\partial f_2(x)$$

where the equality holds if functions f_1, f_2 are regular and $f_1(x) \ge 0$, $f_2(x) \ge 0$. Function $\varphi = f_1 f_2$ is subdifferentially regular in the later case.

Proof. Let us define function $g : R^2 \to R$ by the formula $g(u_1, u_2) = u_1 u_2$. This function is continuously differentiable and, by Theorem 3.3, Lipschitz continuous in a neighborhood of an arbitrary point $u \in R^2$. Moreover, we can write

$$\nabla g(u) = \begin{bmatrix} u_2 \\ u_1 \end{bmatrix} = Pu.$$

By Theorem 3.12, function $g \circ f = f_1 f_2$ is Lipschitz continuous in a neighborhood of point x and one has

$$\partial(f_1 f_2) = \left\{J^T \nabla g(f(x)) : J \in \partial f(x)\right\} = (\partial f(x))^T P f.$$

Formula $\partial(f_1 f_2) \subset \partial f_1(x) f_2(x) + f_1(x)\partial f_2(x)$ and conditions for the equality follow from Corollary 3.1 (c) immediately. $\diamond$

Corollary 3.3 *Let mapping $f : R^n \to R^m$ be Lipschitz continuous in a neighborhood of point $x \in R^n$. Then function $\varphi = (1/2)f^T f$ is Lipschitz continuous in a neighborhood of point x and one has*

$$\partial\varphi(x) = \frac{1}{2}\partial(f^T f)(x) = (\partial(f(x))^T f(x) = \left\{ J^T f(x) : J \in \partial(f(x)) \right\}. \tag{3.20}$$

Proof. Let us define function $g : R^m \to R$ by the formula

$$g(u) = \frac{1}{2}u^T u = \frac{1}{2}\sum_{i=1}^{m} u_i^2.$$

This function is continuously differentiable and, by Theorem 3.3, Lipschitz continuous in a neighborhood of an arbitrary point $u \in R^m$. Moreover, one has $\nabla g(u) = u$. By Theorem 3.12 (a), function $\varphi = g \circ f = (1/2)f^T f$ is Lipschitz continuous in a neighborhood of point x and $\partial\varphi(x) = (\partial f(x))^T \nabla g(f(x)) = (\partial f(x))^T f(x)$ holds. $\diamond$

Theorem 3.13 *Let functions $f_i : R^n \to R$, $1 \le i \le m$, be Lipschitz continuous in a neighborhood of point $x \in R^n$. Then function*

$$\varphi(x) = \max_{1\le i\le m} f_i(x)$$

is Lipschitz continuous in a neighborhood of point x and one has

$$\partial\varphi(x) \subset \text{conv}\ \{\partial f_i(x) : i \in I(x)\}, \tag{3.21}$$

where $I(x) = \{i \in \{1, \dots, m\} : f_i(x) = \varphi(x)\}$. If functions f_i, $1 \le i \le m$, are subdifferentially regular at point x, then function φ is subdifferentially regular at point x and (3.21) holds with equality.

Proof. Let us define function $g : R^m \to R$ by the formula $g(u) = \max(u_1, \dots, u_m)$. This function is convex in R^m, since

$$\begin{aligned} g(\lambda u + (1-\lambda)v) &= \max_{1\le i\le m} (\lambda u_i + (1-\lambda)v_i) \le \lambda \max_{1\le i\le m} (u_i) + (1-\lambda) \max_{1\le i\le m} (v_i) \\ &= \lambda g(u) + (1-\lambda)g(v) \end{aligned}$$

for $u \in R^m$, $v \in R^m$ and $1 \le \lambda \le 1$ so that it is locally Lipschitz by Theorem 2.11. Let $I(u) = \{i \in \{1, \dots, m\} : u_i = g(u)\}$. Then we can write

$$\begin{aligned} g'(u,d) &= \lim_{t\downarrow 0} \frac{g(u+td) - g(u)}{t} = \lim_{t\downarrow 0} \max_{1\le i\le m} \left(\frac{u_i + td_i - g(u)}{t} \right) = \\ &= \lim_{t\downarrow 0} \max_{i\in I(u)} \left(\frac{u_i + td_i - g(u)}{t} \right) = \max_{i\in I(u)} (d_i) \end{aligned}$$

so that $g^0(u,d) = g'(u,d) = \max_{i\in I(u)}(d_i)$ and Definition 3.3 implies

$$\partial g(u) = \left\{ v \in R^n : \max_{i\in I(u)} (d_i) \ge v^T d \ \ \forall d \in R^n \right\}.$$

Let e_i be the i-th column of the unit matrix and $\delta > 0$. If $v_i \neq 0$ for $i \notin I(u)$, the choice $d_i = v_i e_i$ implies the inequality $v^T d = v_i^2 > 0 = \max\{d_i, i \in I(u)\}$, so that $v \notin \partial g(u)$. If $v_i < 0$ for $i \in I(u)$, the choice $d_i = -\delta e_i$ implies the inequality $v^T d = -\delta v_i > -\delta = \max\{d_i, i \in I(u)),$ so that $v \notin \partial g(u)$. If $v_i \geq 0$, $\forall i \in I(u)$ and $\sum_{i\in I(u)} v_i > 1$, the choice $d = \sum_{i\in I(u)} \delta e_i$ implies the inequality $v^T d = \delta \sum_{i\in I(u)} v_i > \delta = \max\{d_i, i \in I(u)\}$, so that $v \notin \partial g(u)$. If $v_i \geq 0$ $\forall i \in I(u)$ and $\sum_{i\in I(u)} v_i < 1$, the choice $d = -\sum_{i\in I(u)} \delta e_i$ implies the inequality $v^T d = -\delta \sum_{i\in I(u)} v_i > -\delta = \max\{d_i, i \in I(u)\}$, so that $v \notin \partial g(u)$. Consequently, the relation

$$\partial g(u) = \left\{ v \in R^n : v_i \geq 0, \sum_{i\in I(u)} v_i = 1, \sum_{i \notin I(u)} v_i = 0 \right\}$$

has to hold. By Corollary 3.1 one has

$$\begin{aligned}
\partial f(x) \quad &\subset \quad \text{conv} \left\{ \sum_{i=1}^{m} v_i u_i : u_i \in \partial f_i(x), v \in \partial g(f(x)) \right\} \\
&= \quad \text{conv} \left\{ \sum_{i\in I(u)} v_i \partial f_i(x) : v_i \geq 0, \sum_{i\in I(u)} v_i = 1 \right\} \\
&= \quad \text{conv} \ \{\partial f_i(x), i \in I(u)\}
\end{aligned}$$

Function g is convex so that it is subdifferentially regular by Theorem 3.5. If functions f_i, $1 \leq i \leq m$, are subdifferentially regular, then function φ is also subdifferentially regular by Theorem 3.5 and since $v_i \geq 0$, $1 \leq i \leq m$, (3.20) holds with the equality. $\diamond$

4 Nonsmooth equations

4.1 Semismooth mappings

Definition 4.1 *Let mapping $f : R^n \to R^m$ be Lipschitz continuous in a neighborhood of point $x \in R^n$. If, for any $h \in R^n$, the limit*

$$\lim_{\substack{J\in\partial f(x+th') \\ h' \to h, t\downarrow 0}} Jh' \tag{4.1}$$

exists (independently of the choice of $J \in \partial f(x+th')$), then mapping f is said to be semismooth at point x. If, for any $h \in R^n$, the limit

$$\lim_{\substack{J\in\partial f(x+th) \\ t\downarrow 0}} Jh \tag{4.2}$$

exists (independently of the choice of $J \in \partial f(x+th)$), then mapping f is said to be weakly semismooth at point x.

Remark 4.1 Since $\partial f(x)$ is a point-to-set mapping, it could seem that the existence of limit (4.1) is exceptional. However, we show in this section (Remark 4.4) that the semismoothness is a property of most interesting locally Lipschitz mappings.

Remark 4.2 Definition 4.1 implies that every semismooth mapping is also weakly semismooth. However, weak semismoothness is not preserved by a composition of mappings and also Theorem 4.6 requires validity of relation (4.1).

Theorem 4.1 *Let mapping $f : R^n \to R^m$ be weakly semismooth at point $x \in R^n$. Then:*

(a) Directional derivative

$$f'(x, h) = \lim_{t \downarrow 0} \frac{f(x + th) - f(x)}{t}$$

exists for an arbitrary vector $h \in R^n$ and one has

$$f'(x, h) = \lim_{\substack{J \in \partial f(x+th) \\ t \downarrow 0}} Jh.$$

(b) Mapping $f'(x, .) : R^n \to R^m$ is positive homogeneous and Lipschitz continuous. (This assertion does not require semismoothness of mapping f. Lipschitz continuity and existence of directional derivative $f'(x, .)$ suffice.)

Proof (a) Choose vector $h \in R^n$ and sequence $t_i \downarrow 0$ arbitrarily. Since mapping f is Lipschitz continuous in a neighborhood of point x, we can suppose without loss of generality that it is Lipschitz continuous in all intervals $[x, x + t_i h]$. Using Theorem 3.11 we obtain

$$\frac{f(x + t_i h) - f(x)}{t_i} \in \partial f\left([x, x + t_i h]\right) h.$$

By Theorem 2.3, there are at most $m + 1$ elements $J_i^k \in \partial f(x + t_i^k h)$, $t_i^k \in [0, t_i]$, $1 \le k \le m + 1$, so that

$$\frac{f(x + t_i h) - f(x)}{t_i} = \sum_{k=1}^{m+1} \lambda_i^k J_i^k h,$$

where $0 \le \lambda_i^k \le 1$ and $\sum_{i=1}^{m+1} \lambda_i^k = 1$. Since the interval $[0, 1]$ is compact, we can assume (by passing to a subsequence, if necessary) that $\lambda_i^k \to \lambda^k$, $1 \le k \le m + 1$. Then (4.2) implies

$$\begin{aligned}
\lim_{i \to \infty} \frac{f(x + t_i h) - f(x)}{t_i} &= \lim_{i \to \infty} \left(\sum_{k=1}^{m+1} \lambda_i^k J_i^k h \right) = \sum_{k=1}^{m+1} \left(\lim_{i \to \infty} \lambda_i^k \right) \left(\lim_{i \to \infty} J_i^k h \right) \\
&= \left(\sum_{k=1}^{m+1} \lambda^k \right) \lim_{\substack{J \in \partial f(x+th) \\ t \downarrow 0}} Jh = \lim_{\substack{J \in \partial f(x+th) \\ t \downarrow 0}} Jh
\end{aligned}$$

so that the limit on the left hand side does not depend on the choice of $t_i \downarrow 0$ and it is equal to the directional derivative $f'(x, h)$. $\diamond$

Remark 4.3 If mapping $f : R^n \to R$ is semismooth, then the Hadamard directional derivative exists and

$$\lim_{\substack{h' \to h \\ t \downarrow 0}} \frac{f(x + th') - f(x)}{t} = \lim_{\substack{J \in \partial f(x+th') \\ h' \to h,\, t \downarrow 0}} Jh'$$

holds (the expression on the left hand side is the Hadamard directional derivative). If the Hadamard directional derivative exists, it is equal to ordinary directional derivative $f'(x, h)$.

Theorem 4.2 *If functions $f_i : R^n \to R$, $1 \le i \le m$, are semismooth at a point $x \in R^n$, then the mapping $f = [f_1, \dots, f_m]^T$ is semismooth at point x as well.*

Proof. By assumption, the limits

$$\lim_{\substack{g_i \in \partial f_i(x+th') \\ h' \to 0,\, t \downarrow 0}} g_i^T h',$$

$1 \le i \le m$ exist. By Theorem 3.10 (a), one has $g_i = J^T e_i \in \partial f_i(x + th')$, $1 \le i \le m$, for any element $J \subset \partial f(x + th')$ (e_i is i-th column of the unit matrix of order m) so that limit (4.1) exists. $\diamond$

Theorem 4.3 *If function $f : R^n \to R$ is continuously differentiable in a neighborhood of point $x \in R^n$, then it is semismooth at point x.*

Proof. For continuously differentiable functions, one has

$$\lim_{\substack{g \in \partial f(x+th') \\ h' \to h,\, t \downarrow 0}} g^T h' = \lim_{\substack{h' \to h \\ t \downarrow 0}} \left(\nabla f(x + th')\right)^T h' = (\nabla f(x))^T h.$$

$\diamond$

Theorem 4.4 *If function $f : R^n \to R$ is convex in a neighborhood of point $x \in R^n$, it is semismooth at point x.*

Proof. Let function f be convex in $B(x, \varepsilon)$, $x + th' \in B(x, \varepsilon)$ and $g \in \partial f(x + th')$. Then Theorem 2.13 (d) implies

$$f(x) - f(x + th') \ge g^T(x - (x + th'))$$

or

$$\frac{f(x + th') - f(x)}{t} \le g^T h'.$$

On the other hand, Definition 2.7 gives

$$g^T h' \le f'(x + th', h').$$

Using these inequalities together with Definition 2.6 and Theorem 2.12 (c), we obtain

$$\begin{aligned} f'(x, h) &\le \limsup_{\substack{h' \to h \\ t \downarrow 0}} \frac{f(x + th') - f(x)}{t} \le \liminf_{\substack{g \in \partial f(x+th') \\ h' \to h,\, t \downarrow 0}} g^T h' \le \limsup_{\substack{g \in \partial f(x+th') \\ h' \to h,\, t \downarrow 0}} g^T h' \\ &\le \limsup_{h' \to h,\, t \downarrow 0} f'(x + th', h') \le f'(x, h), \end{aligned}$$

which proves the existence of the required limit. $\diamond$

Theorem 4.5 *Let mapping $f : R^n \to R^m$ be semismooth at point $x \in R^n$ and function $g : R^m \to R$ be semismooth at the point $f(x)$. Then the composite mapping $\varphi = g \circ f$ is semismooth at point x.*

Proof. (a) Let $h \in R^n$. Since mapping f is Lipschitz continuous in a neighborhood of point x (with some modulus L), one has

$$\lim_{h' \to h,\, t \downarrow 0} \frac{\|f(x+th') - f(x+th)\|}{t} \le \lim_{h' \to h} L\|h' - h\| = 0.$$

Therefore, we can write

$$\begin{aligned}
\lim_{h' \to h,\, t \downarrow 0} \frac{f(x+th') - f(x)}{t} &= \lim_{t \downarrow 0} \frac{f(x+th) - f(x)}{t} \\
&+ \lim_{h' \to h,\, t \downarrow 0} \frac{f(x+th') - f(x+th)}{t} = f'(x,h)
\end{aligned}$$

so that $f(x+th') = f(x) + tf'$, where $f' \to f'(x,h)$, whenever $h' \to h$ a $t \downarrow 0$.
(b) Let $h \in R^n$. Let $x_k = x + t_k h_k$, where $h_k \to h$ and $t_k \downarrow 0$. By Theorem 3.12, one has $\partial\varphi(x_k) \subset S_k$, where $S_k \subset R^n$ is a compact set on the right hand side of (3.14) (with x_k instead of x). Let

$$\begin{aligned}
w_k^- &= (J_k^-)^T v_k^- = \arg\min_{w \in S_k} w^T h, \quad v_k^- \in \partial g(f(x_k)), \quad J_k^- \in \partial f(x_k), \\
w_k^+ &= (J_k^+)^T v_k^+ = \arg\max_{w \in S_k} w^T h, \quad v_k^+ \in \partial g(f(x_k)), \quad J_k^+ \in \partial f(x_k).
\end{aligned}$$

Then

$$(w_k^-)^T h \le w_k^T h \le (w_k^+)^T h \tag{4.3}$$

holds for an arbitrary vector $w_k \in \partial\varphi(x_k) \subset S_k$. Since all values in these formulas are bounded by Theorem 3.2 (a), we can assume (passing to subsequences, if necessary) that

$$\begin{aligned}
J_k^- &\to J^- \in \partial f(x), \qquad v_k^- \to v^- \in \partial g(f(x)), \\
J_k^+ &\to J^+ \in \partial f(x), \qquad v_k^+ \to v^+ \in \partial g(f(x))
\end{aligned}$$

(we use Theorem 3.2 (c)). Since the mapping f is semismooth, one has $J^- h = J^+ h = f'(x,h)$ so that, using (4.3), we obtain

$$(v^-)^T f'(x,h) \le \liminf_{k \to \infty} w_k^T h \le \limsup_{k \to \infty} w_k^T h \le (v^+)^T f'(x,h).$$

Since function g is semismooth and (a) implies $f(x_k) = f(x + t_k h_k) = f(x) + t_k f_k'$, where $f_k' \to f'(x,h)$, whenever $h_k \to h$ and $t_k \downarrow 0$, we can use Definition 4.1 which gives

$$(v^-)^T f'(x,h) = \lim_{k \to \infty} (v_k^-)^T f'(x,h) = \lim_{k \to \infty} (v_k^+)^T f'(x,h) = (v^+)^T f'(x,h).$$

This proves the existence of a limit of sequence $w_k^T h$, which is independent of the choice of vector $w_k \in \partial\varphi(x_k)$. $\diamond$

Corollary 4.1 *Let the mapping $f : R^n \to R^m$ be semismooth at point $x \in R^n$ and let function $g : R^n \to R$ be either continuously differentiable or convex in a neighborhood of point $f(x)$. Then function $\varphi = g \circ f$ is semismooth at point x.*

Proof. This assertion follows immediately from Theorem 4.3, Theorem 4.4 and Theorem 4.5. $\diamond$

Corollary 4.2 *Let functions $f_i : R^n \to R$, $1 \le i \le m$ be semismooth at point $x \in R^n$ and $\lambda_i \in R$, $1 \le i \le m$. Then functions $\varphi_1 = \sum_{i=1}^m \lambda_i f_i$ (linear combination) and $\varphi_2 = \prod_{i=1}^m f_i$ (product) are semismooth at point x.*

Proof. By Theorem 4.2, the mapping $f = [f_1, \dots, f_m]^T$ is semismooth at point x. Functions $g_1(u) = \sum_{i=1}^m \lambda_i u_i$ and $g_2(u) = \prod_{i=1}^m u_i$ are continuously differentiable so that, by Corollary 4.1, functions $\varphi_1 = g_1 \circ f$ and $\varphi_2 = g_2 \circ f$ are semismooth at point x. $\diamond$

Corollary 4.3 *Let functions $f_i : R^n \to R$, $1 \le i \le m$ be semismooth at a point $x \in R^n$. Then function $\varphi = \|[f_1, \dots, f_m]^T\|$, where $\|\cdot\|$ is an arbitrary norm in R^m, is semismooth at point x. Particularly, functions $\varphi_1 = \max_{1 \le i \le m}(|f_i|)$ (maximum of absolute values) and $\varphi_2 = \sum_{i=1}^m |f_i|$ (sum of absolute values) are semismooth at point x. Furthermore, function $\varphi_3 = \max_{1 \le i \le m}(f_i)$ (poinwise maximum) is semismooth at point x.*

Proof. By Theorem 4.2, the mapping $f = [f_1, \dots, f_m]^T$ is semismooth at point x. Function $g(u) = \|u\|$ is convex, since properties of vector norm imply that, for $0 \le \lambda \le 1$, one has

$$g(\lambda u + (1-\lambda)v) = \|\lambda u + (1-\lambda)v\| \le \lambda\|u\| + (1-\lambda)\|v\|.$$

Therefore, function $\varphi = g \circ f$ is semismooth by Corollary 4.1. Function $g(u) = \max_{1 \le i \le m}(u_i)$ is also convex (proof of Theorem 3.13) so that function φ_3 is semismooth by Corollary 4.1 $\diamond$

Corollary 4.4 *(Inverse of Theorem 4.2). Let mapping $f : R^n \to R^m$ be semismooth at point $x \in R^n$, where $f = [f_1, \dots, f_m]^T$. Then functions $f_i : R^n \to R$, $1 \le i \le m$, are semismooth at point x.*

Proof. Obviously, $f_i = g_i \circ f$, $1 \le i \le m$, where functions $g_i : R^m \to R$, defined by formulas $g_i(u) = u_i$, are continuously differentiable. Therefore, semismoothness of functions $f_i : R^n \to R$, $1 \le i \le m$, follows from Corollary 4.1. $\diamond$

Corollary 4.5 *A linear combination of semismooth mappings is a semismooth mapping. The dot product of semismooth mappings is a semismooth function.*

Proof. By Corollary 4.4, the elements of semismooth mappings are semismooth functions. By Corollary 4.2, a linear combination of semismooth functions is a semismooth function so that a linear combination of semismooth mappings is a semismooth mapping by Theorem 4.2. Semismoothness of dot product follows from Theorem 4.2, Corollary 4.4 and Corollary 4.2. $\diamond$

Remark 4.4 From the above text, we can deduce that, starting from continuously differentiable and convex mappings, we always get semismooth mappings by the usual operations (sum, product, maximum, composition). For this reason, the teory of semismooth mappings has broad applications in practical problems. Moreover, semismoothness is a basic assumption for the construction of numerical metods for solving nonsmooth equations.

Theorem 4.6 *Let the mapping $f : R^n \to R^m$ be Lipschitz continuous in a neighborhood of point $x \in R^n$. Then f is semismooth at point x if and only if there exists directional derivative $f'(x, h)$ and*

$$Jh - f'(x, h) = o(\|h\|), \tag{4.4}$$

holds whenever $h \to 0$ and $J \in \partial f(x + h)$.

Proof. (a) Let mapping $f : R^n \to R^m$ be semismooth. Let $h_i \to 0$ and $J_i \in \partial f(x + h_i)$. We have to prove that

$$\lim_{i\to\infty} \frac{J_i h_i - f'(x, h_i)}{\|h_i\|} = \lim_{i\to\infty} \left(J_i \frac{h_i}{\|h_i\|} - f'\left(x, \frac{h_i}{\|h_i\|}\right)\right) = 0.$$

Assume, for contradiction purposes, that there are a subsequence $\{h_i'\} \subset \{h_i/\|h_i\|\}$, a corresponding subsequence $\{t_i\} \subset \{\|h_i\|\}$ and a number $\varepsilon > 0$ such that

$$J_i h_i' - f'(x, h_i') \geq \varepsilon \quad \forall i, \tag{4.5}$$

where $J_i \in \partial f(x + h_i)$. Since $\|h_i'\| = 1\ \forall i$, we can choose this subsequence in such a way that $h_i' \to h$. Then Remark 4.3 implies

$$\lim_{i\to\infty} J_i h_i' = f'(x, h),$$

which is in contradiction to (4.5), since function $f'(x, .)$ is continuous by Theorem 4.1 (b).
(b) Assume now that mapping $f : R^n \to R^m$ is not semismooth. Then there are vector $h \in R^n$ (we assume without loss of generality that $\|h\| = 1$), sequences $h_i' \to h$, $t_i \downarrow 0$, $J_i \in \partial f(x + t_i h_i')$ and a number $\varepsilon > 0$ such that

$$\|J_i h_i' - f'(x, h)\| \geq 2\varepsilon \quad \forall i \tag{4.6}$$

(limit (4.1) would exist equal to $f'(x, h)$ in the opposite case). Since the directional derivative is Lipschitz continuous by Theorem 4.1 (b), one has $\|f'(x, h_i') - f'(x, h)\| \leq \varepsilon$ for sufficiently large indices, which together with (4.6) gives

$$\begin{aligned} \|J_i h_i' - f'(x, h_i')\| &= \|J_i h_i' - f'(x, h) - (f'(x, h_i') - f'(x, h))\| \\ &\geq \|J_i h_i' - f'(x, h)\| - \|(f'(x, h_i') - f'(x, h))\| \geq \varepsilon. \end{aligned}$$

Set $h_i = t_i h_i'$. Since $\|h_i'\| \to 1$ and $t_i \to 0$, one has $\|h_i\| \to 0$, but the above inequality implies

$$\liminf_{i\to\infty} \frac{\|J_i h_i - f'(x, h_i)\|}{\|h_i\|} = \liminf_{i\to\infty} \frac{\|J_i h_i' - f'(x, h_i')\|}{\|h_i'\|} = \liminf_{i\to\infty} \|J_i h_i' - f'(x, h_i')\| \geq \varepsilon > 0,$$

which is in contradiction to (4.4). ◇

Remark 4.5 Considering Theorem 4.6, a semismooth mapping is frequently defined to be a locally Lipschitz mapping, which satisfies condition (4.4).

Definition 4.2 *Mapping $f : R^n \to R^m$ is said to be differentiable in the Bouligand sense (B-differentiable) at point $x \in R^n$, if there is a positively homogeneous mapping $f'(x, h)$ (directional derivative) such that*

$$f(x + h) - f(x) - f'(x, h) = o\left(\|h\|\right), \tag{4.7}$$

whenever $h \to 0$.

Theorem 4.7 *A semismooth mapping is B-differentiable.*

Proof. Let the mapping $f : R^n \to R^m$ be semismooth at point $x \in R^n$. Let $h_i \to 0$. We have to prove that

$$\lim_{i\to\infty} \frac{f(x+h_i) - f(x) - f'(x,h_i)}{\|h_i\|} = \lim_{i\to\infty} \left(\frac{f(x+h_i) - f(x)}{\|h_i\|} - f'\left(x, \frac{h_i}{\|h_i\|}\right)\right) = 0.$$

Assume, for contradiction purposes, that there are a subsequence $\{h_i'\} \subset \{h_i/\|h_i\|\}$, a corresponding subsequence $\{t_i\} \subset \{\|h_i\|\}$ and a number $\varepsilon > 0$ such that

$$\left| \frac{f(x+t_i h_i') - f(x)}{t_i} - f'(x,h_i') \right| \geq \varepsilon \quad \forall i. \tag{4.8}$$

Since $\|h_i'\| = 1 \ \forall i$, we can choose this subsequence in such a way that $h_i' \to h$. Since mapping f is semismooth, the Hadamard directional derivative exists by Remark 4.3 which is identical to the usual directional derivative. Therefore, one has

$$\lim_{i\to\infty} \frac{f(x+t_i h_i') - f(x)}{t_i} = f'(x,h),$$

which is in contradiction to (4.8), since function $f'(x,\cdot)$ is continuous by Theorem 4.1 (b). $\diamond$

Corollary 4.6 *Let mapping $f : R^n \to R^m$ be semismooth at point $x \in R^n$. Then one has*

$$f(x+h) - f(x) - Jh = o\left(\|h\|\right), \tag{4.9}$$

whenever $h \to 0$ and $J \in \partial f(x+h)$.

Proof. The assertion follows immediately from Theorem 4.6 and Theorem 4.7. $\diamond$

4.2 Solution of nonsmooth equations

Now we will consider solving a system of equations

$$f(x) = 0, \tag{4.10}$$

where $f : R^n \to R^n$ is a semismooth mapping. First we will investigate the inexact Newton method, which is iterative and generates a sequence $\{x_k\}$ by the formula

$$x_{k+1} = x_k + d_k, \tag{4.11}$$

where vector d_k is chosen in such a way that

$$\omega_k = \frac{\|A_k d_k + f(x_k)\|}{\|f(x_k)\|} \leq \omega \tag{4.12}$$

and matrix A_k is chosen in such a way that

$$\Delta_k = \|A_k - J_k\| \leq \Delta \tag{4.13}$$

for some element $J_k \in \partial_B f(x_k)$. At the same time, $\omega \geq 0$, $\Delta \geq 0$ and norms in (4.12) and (4.13) are Euclidean.

Definition 4.3 *The locally Lipschitz mapping $f : R^n \to R^n$ is said to be strongly BD-regular at point $x \in R^n$, if all matrices $J \in \partial_B f(x)$ are nonsingular (set $\partial_B f(x)$ is introduced in Definition 3.5).*

Remark 4.6 In the iterative process (4.11)-(4.13), we assume that A_k approximates an element from $\partial_B f(x_k)$, since the nonsingularity of all elements of $\partial_B f(x_k) \subset \partial f(x_k)$ can be assured much easier then the nonsingularity of all elements of $\partial f(x_k)$.

Theorem 4.8 *Let the locally Lipschitz mapping $f : R^n \to R^n$ be strongly BD-regular at point $x \in R^n$. Then there is a number $\delta > 0$ and a constant $c \geq 0$ such that all matrices $J \in \partial_B f(y)$ are nonsingular and $\|J^{-1}\| \leq c$ holds whenever $y \in B(x, \delta)$.*

Proof. First we prove existence of a number $\delta > 0$ and a constant $c \geq 0$ such that all Jacobian matrices $\mathcal{J} f(z)$ are nonsingular and

$$\|(\mathcal{J} f(z))^{-1}\| \leq c \tag{4.14}$$

holds whenever $z \in B(x, \delta) \backslash \Omega_f$. (set Ω_f is introduced in Definition 3.5). Assume, for contradiction purposes, that (4.14) does not hold. Then a sequence $x_i \to x$, $x_i \in B(x, \delta) \backslash \Omega_f$ exists such that either all the Jacobian matrices $\mathcal{J} f(x_i)$ are singular or $\|(\mathcal{J} f(x_i))^{-1}\| \to \infty$. Since mapping f is Lipschitz continuous in a neighborhood of point x, the Jacobian matrices $\mathcal{J} f(x_i)$ are bounded in a neighborhood of point x. Therefore, there is a subsequence $\{x_i'\} \subset \{x_i\}$ such that $\mathcal{J} f(x_i') \to J$. Continuous dependence of eigenvalues on matrix coefficients implies that J has to be singular. By Definition 3.5, $J \in \partial_B f(x)$ holds, which is in contradiction to Definition 4.3. Let now $y \in B(x, \delta) \cap \Omega_f$ and $J \in \partial_B f(y)$. Then there is a number $0 < \delta' < \delta$ such that $B(y, \delta') \subset B(x, \delta)$ and (4.14) holds whenever $z \in B(y, \delta') \backslash \Omega_f$. Since Definition 3.5 implies

$$J = \lim_{i \to \infty} \mathcal{J} f(y_i)$$

for some sequence $y_i \to y$, $y_i \in B(y, \delta') \backslash \Omega_f$, we obtain inequality $\|J^{-1}\| \leq c$ from (4.14) and from continuous dependence of eigenvalues on matrix coefficients. $\diamond$

Theorem 4.9 *Let mapping $f : R^n \to R^n$ be semismooth and strongly BD-regular at point $x^* \in R^n$ such that $f(x^*) = 0$. Then there are numbers $\varepsilon > 0$, $\omega > 0$ and $\Delta > 0$ such that if $x_1 \in B(x^*, \varepsilon)$, then the iterative process (4.11)-(4.13) is well defined (matrices A_k are nonsingular) and sequence $\{x_k\}$ converges to point x^* Q-linearly. If, in addition, $\omega_k \to 0$ and $\Delta_k \to 0$ hold, then sequence $\{x_k\}$ converges to point x^* Q-superlinearly and also sequence $\{f(x_k)\}$ converges to zero Q-superlinearly.*

Proof. Let c and δ be the numbers used in Theorem 4.8. Let us set $\Delta = 1/(5c)$ and choose $\varepsilon \leq \delta$ in such a way that mapping f is Lipschitz continuous (with some modulus L) in $B(x^*, \varepsilon)$ and

$$\|f(x) - f(x^*) - J(x - x^*)\| \leq \frac{\Delta}{2} \|x - x^*\| \quad \forall J \in \partial_B f(x) \tag{4.15}$$

holds whenever $x \in B(x^*, \varepsilon)$ (this is possible by (4.9)). Furthermore, set $\omega = \Delta/(2L)$. Assume that $x_k \in B(x^*, \varepsilon)$ (this holds for $k = 1$). Then Theorem 4.8 implies $\|J_k^{-1}\| \leq c$. Obviously

$$A_k^{-1} + J_k^{-1}(A_k - J_k)A_k^{-1} = J_k^{-1}.$$

Since difference of norms is not greater than a norm of difference, we can write

$$\|A_k^{-1}\| - \|J_k^{-1}\|\|A_k - J_k\|\|A_k^{-1}\| \le \|J_k^{-1}\|$$

or

$$\|A_k^{-1}\| \le \frac{\|J_k^{-1}\|}{1 - \|J_k^{-1}\|\|A_k - J_k\|} \le \frac{c}{1 - c\Delta} = \frac{5}{4}c,$$

which by (4.11)-(4.13) and (4.15) (together with $f(x^*) = 0$) gives

$$\begin{aligned}
\|x_{k+1} - x^*\| &= \|x_k + d_k - x^*\| = \|x_k + A_k^{-1}(A_k d_k + f(x_k) - f(x_k)) - x^*\| \\
&= \|A_k^{-1}(A_k d_k + f(x_k) - (f(x_k) - J_k(x_k - x^*)) + (A_k - J_k)(x_k - x^*))\| \\
&\le \|A_k^{-1}\| \left(\|f(x_k) - f(x^*) - J_k(x_k - x^*)\|\right. \\
&\quad \left. + \|A_k - J_k\|\|x_k - x^*\| + \omega_k\|f(x_k) - f(x^*)\|\right) \\
&\le \frac{5}{4}c\left(\|f(x_k) - f(x^*) - J_k(x_k - x^*)\|\right. \\
&\quad \left. + \Delta_k\|x_k - x^*\| + \omega_k L\|x_k - x^*\|\right) \\
&\le \frac{5}{4}c\left(\frac{1}{2}\Delta + \Delta + \frac{1}{2}\Delta\right)\|x_k - x^*\| = \frac{1}{2}\|x_k - x^*\|. \qquad (4.16)
\end{aligned}$$

This implies $x_{k+1} \in B(x^*, \varepsilon)$ so that we can continue in the same way. Therefore, we have proved by induction that $x_{k+1} \in B(x^*, \varepsilon)$ and $\|x_{k+1} - x^*\| \le (1/2)\|x_k - x^*\|$ hold in all iterations or that sequence $\{x_k\}$ converges to point x^* Q-linearly. Assume now that $\omega_k \to 0$ and $\Delta_k \to 0$. Then (4.9) and (4.16) imply

$$\begin{aligned}
\|x_{k+1} - x^*\| &\le \frac{5}{4}c\left(\|f(x_k) - f(x^*) - J_k(x_k - x^*)\|\right. \\
&\quad \left. + \Delta_k\|x_k - x^*\| + \omega_k L\|x_k - x^*\|\right) \\
&= \frac{5}{4}c\left(o(\|x_k - x^*\|) + o(\|x_k - x^*\|) + o(\|x_k - x^*\|)\right) \\
&= o(\|x_k - x^*\|) \qquad (4.17)
\end{aligned}$$

and sequence $\{x_k\}$ converges to point x^* Q-superlinearly. Since $f(x^*) = 0$, we can write

$$\lim_{k\to\infty} \frac{\|f(x_{k+1})\|}{\|x_k - x^*\|} \le L \lim_{k\to\infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = 0 \qquad (4.18)$$

by (4.17). Using (4.11)-(4.13) a (4.16) we obtain

$$\begin{aligned}
\|x_k - x^*\| &\le \|x_{k+1} - x_k\| + \|x_{k+1} - x^*\| \\
&\le \|A_k^{-1}\|\|A_k d_k + f(x_k)\| + \|A_k^{-1}\|\|f(x_k)\| + \|x_{k+1} - x^*\| \\
&\le \frac{5}{4}c(1 + \omega)\|f(x_k)\| + \frac{1}{2}\|x_k - x^*\|
\end{aligned}$$

or

$$\|x_k - x^*\| \le \frac{5}{2}c(1 + \omega)\|f(x_k)\|$$

so that (4.18) implies

$$\lim_{k\to\infty} \frac{\|f(x_{k+1})\|}{\|f(x_k)\|} \le \frac{5}{2} c(1+\omega) \lim_{k\to\infty} \frac{\|f(x_{k+1})\|}{\|x_k - x^*\|} = 0$$

and $\{f(x_k)\}$ converges to zero Q-superlinearly. $\diamond$

Theorem 4.9 says that the inexact Newton metod (4.11)-(4.13) is locally convergent or that it converges whenever the starting point $x_1 \in R^n$ is sufficiently close to the solution $x^* \in R^n$. For ensuring the global convergence (convergence from an arbitrary point) it is necessary to replace the relation (4.11) by a line search. In the following algorithm, we use the line search function

$$\varphi(x) = \frac{1}{2} f^T(x) f(x)$$

and matrices A_k are chosen in such a way that $A_k = J_k$ (so that $\Delta_k = 0$).

Algorithm 4.1

Data $\varrho, \sigma \in (0,1)$, $\omega \in (0, 1-\sigma)$, $\varepsilon > 0$.

Step 1 (Initiation). Choose starting point $x_1 \in R^n$ and set $k = 1$.

Step 2 (Direction vector). If $\varphi(x) \le \varepsilon$, terminate the computation. In the opposite case, choose $J_k \in \partial_B f(x_k)$ and determine a direction vector d_k satisfying condition

$$\omega_k = \frac{\|J_k d_k + f(x_k)\|}{\|f(x_k)\|} \le \omega. \tag{4.19}$$

Step 3 (Line search). Let $t_k = \varrho^{i_k}$, where i_k is the smallest nonnegative integer satisfying condition

$$\varphi(x_k + \varrho^i d_k) - \varphi(x_k) \le -2\sigma \varrho^i \varphi(x_k). \tag{4.20}$$

Step 4 (Update). Set $x_{k+1} := x_k + t_k d_k$ and $k := k+1$. Go to Step 2.

Theorem 4.10 *Let the set $X = \{x \in R^n : \varphi(x) \le \varphi(x_1)\}$ be compact, let the mapping $f : R^n \to R$ be semismooth and strongly BD-regular on $X \subset R^n$ and the function $\varphi(x)$ be continuously differentiable on $X \subset R^n$. Then:*

(a) Any cluster point of sequence $\{x_k\}$ generated by Algorithm 4.1 is a solution of equation (4.10).

(b) If $\sigma < 1/2$ and $\omega_k \to 0$, then $x_k \to x^*$ superlinearly.

Proof. (a) Since f is strongly BD-regular on $X \subset R^n$ and set X is compact, there is a constant $c > 0$ such that $\|J_k^{-1}\| \le c$ holds in every iteration. Therefore, Step 2 of Algorithm 4.1 is well defined and (4.19) implies

$$\|d_k\| = \|J_k^{-1}(J_k d_k + f(x_k)) - J_k^{-1} f(x_k)\| \le (1+\omega)\|J_k^{-1}\|\|f_k\| \le c(1+\omega)\sqrt{2\varphi(x_1)}. \tag{4.21}$$

We show that Step 3 of Algorithm 4.1 is also well defined. Assume, for contradicition purposes, that

$$\varphi(x_k + \varrho^i d_k) - \varphi(x_k) > -2\sigma\varrho^i\varphi(x_k)$$

holds for any exponent i which gives

$$\varphi'(x_k, d_k) \geq -2\sigma\varphi(x_k)$$

by passing to the limit. Since φ is continuously differentiable, Corollary 3.3 and (4.19) imply

$$\begin{aligned}
\varphi'(x_k, d_k) &= (\nabla\varphi(x_k))^T d_k = f^T(x_k)J_k d_k \\
&= f^T(x_k)f(x_k) + f^T(x_k)J_k d_k - f^T(x_k)f(x_k) \\
&\leq \|f(x_k)\|\|f(x_k) + J_k d_k\| - \|f(x_k)\|^2 \\
&\leq (\omega - 1)\|f(x_k)\|^2 = -2(1-\omega)\varphi(x_k).
\end{aligned} \tag{4.22}$$

Since $\varphi(x_k) \neq 0$ (the iterative process terminate in Step 2 in the opposite case), we obtain, by comparing the above inequalities, that $\sigma \geq 1 - \omega$, which is in contradiction to assumption $\sigma < 1 - \omega$. Let $\{x_k\}$ be a sequence generated by Algorithm 4.1. Since $x_k \in X$ and $X \subset R^n$ is compact, there is at least one cluster point $x^* \in X$ of sequence $\{x_k\}$. Therefore, there is an index set K such that $x_k \overset{K}{\to} x^*$. We will investigate two separate cases.

(1) Assume first that $t_k \geq \tau > 0$, $\forall k \in K$. Then (4.20) implies

$$\begin{aligned}
\varphi(x_1) &\geq \varphi(x_1) - \lim_{k\to\infty}\varphi(x_{k+1}) = \sum_{k=1}^{\infty}(\varphi(x_k) - \varphi(x_{k+1})) \\
&\geq \sum_{k=1}^{\infty} 2\sigma t_k\varphi(x_k) \geq 2\tau\sigma\sum_{k\in K}\varphi(x_k)
\end{aligned}$$

so that necessarily $\varphi(x_k) \overset{K}{\to} 0$, which together with $x_k \overset{K}{\to} x^*$ gives $\varphi(x^*) = 0$ (since function φ is continuous).

(2) Assume now that $t_k \overset{K_1}{\to} 0$ for some subset $K_1 \subset K$. This implies $i_k \overset{K_1}{\to} \infty$, so that $i_k > 0$ for sufficiently large indices $k \in K_1$. Since (4.20) does not hold for $i = i_k - 1$, we can use the mean value theorem

$$(\nabla\varphi(x_k'))^T d_k = \frac{\varphi\left(x_k + \frac{t_k}{\varrho}d_k\right) - \varphi(x_k)}{\frac{t_k}{\varrho}} > -2\sigma\varphi(x_k),$$

where $x_k' \in (x_k, x_k + (t_k/\varrho)d_k)$. Since sequence $\{\|d_k\|\}_{K_1}$ is bounded by (4.21), it has at least one cluster point d^*. Therefore, there is a subset $K_2 \subset K_1$ such that $d_k \overset{K_2}{\to} d^*$, which together with $x_k \overset{K_2}{\to} x^*$ and $t_k \overset{K_2}{\to} 0$ (so that $x_k' \overset{K_2}{\to} x^*$) and passing to limit gives

$$(\nabla\varphi(x^*))^T d^* \geq -2\sigma\varphi(x^*).$$

On the other hand (4.22) implies $(\nabla\varphi(x_k))^T d_k \leq -2(1-\omega)\varphi(x_k)$, which, passing to limit, gives

$$(\nabla\varphi(x^*))^T d^* \leq -2(1-\omega)\varphi(x^*).$$

Since $\sigma < 1 - \omega$ by assumption, we obtain $\varphi(x^*) = 0$ by comparing both above inequalities. Thus we have proved that if x^* is a cluster point of a sequence generated by Algorithm 4.1, then $\varphi(x^*) = 0$ and, therefore $f(x^*) = 0$.
(b) Let set K be the index set used in part (a) of this proof. We want to show that $x_{k+1} = x_k + d_k$ holds for sufficiently large indices $k \in K$, which allows us to use the same considerations as in the proof of Theorem 4.9. Since $x_k \overset{K}{\to} x^*$, $\omega_k \overset{K}{\to} 0$ and $\Delta_k = 0$, the assumptions used in the proof of Theorem 4.9 (i.e. $x_k \in B(x^*, \varepsilon)$ and $\omega_k \le 1/(10cL)$) are satisfied for sufficiently large indices $k \in K$ so that the point $x_k + d_k$ satisfies (4.16) (with $x_k + d_k$ instead of x_{k+1}) and one has

$$\lim_{k \overset{K}{\to} \infty} \frac{\|x_k + d_k - x^*\|}{\|x_k - x^*\|} = 0,$$
$$\lim_{k \overset{K}{\to} \infty} \frac{\|f(x_k + d_k)\|}{\|f(x_k)\|} = 0.$$

Since $\sigma < 1/2$, there is an index $\overline{k} \in K$ such that $\|f(x_k + d_k)\| \le (1 - 2\sigma)\|f(x_k)\|$, whenever $k \in K$ a $k \ge \overline{k}$. For such indices, one has

$$\begin{aligned} \frac{\varphi(x_k + d_k) - \varphi(x_k)}{\varphi(x_k)} &= \frac{\left(\|f(x_k + d_k)\| - \|f(x_k)\|\right)\left(\|f(x_k + d_k)\| + \|f(x_k)\|\right)}{\|f(x_k)\|^2} \\ &\le \frac{\|f(x_k + d_k)\| - \|f(x_k)\|}{\|f(x_k)\|} \le -2\sigma \end{aligned}$$

so that condition (4.20) is satisfied with $i_k = 0$. Therefore $x_{k+1} = x_k + d_k$ holds and, using (4.16), we can supply index $k + 1$ into set K formally. If we continue in the same way, we can see (as in the proof of Theorem 4.9), that $x_k \to x^*$ superlinearly. $\diamond$

Remark 4.7 The requirement of continuous differentiability of function $\varphi = (1/2)f^T f$ may seem to be unrealistic at the first look, since mapping f is not continuously differentiable. Surprisingly, this requirement is satisfied in many important applications.

Remark 4.8 In Algorithm 4.1, the matrix $J_k \in \partial_B f(x_k)$ is used. Since mapping f is differentiable almost everywhere by the Rademacher theorem, one usually has $x_k \notin \Omega_f$ so that $J_k = \mathcal{J} f(x_k)$. If $x_k \in \Omega_f$, the computation of $J_k \in \partial_B f(x_k)$ may be more difficult. It follows from Definition 3.5 that

$$\partial_B f(x_k) \subset [\partial_B f_1(x_k), \dots, \partial_B f_n(x_k)]^T \triangleq \partial_b f(x_k),$$

and determination of $\partial_b f(x_k)$ is usually easier than determination of $\partial_B f(x_k)$. Therefore, the question arises if it is possible to choose $J_k \in \partial_b(x_k)$. The answer is positive. Let $J \in \partial_b f(x)$. Since functions $f_1, \dots, f_n$ are semismooth by Corollary 4.4,

$$\begin{gathered} f_1(x + h) - f_1(x) - e_1^T J h = o\left(\|h\|\right), \\ \dots\dots\dots\dots\dots\dots\dots\dots \\ f_n(x + h) - f_1(x) - e_n^T J h = o\left(\|h\|\right), \end{gathered}$$

holds by Corollary 4.6 and n is finite, the key relation (4.9) holds for $J \in \partial_b f(x)$ as well and nothing is changed in the proofs of Theorem 4.9 and Theorem 4.10.

4.3 Application of nonsmooth equations

Definition 4.4 *Let mapping $p : R^n \to R^n$ be continuously differentiable. The nonlinear complementarity problem (NCP) is the problem of finding a point $x^* \in R^n_+$ such that $p(x^*) \in R^n_+$ and $(x^*)^T p(x^*) = 0$, i.e.*

$$x_i^* \geq 0, \quad p_i(x^*) \geq 0, \quad x_i^* p_i(x^*) = 0 \tag{4.23}$$

for any index $1 \leq i \leq n$.

The nonlinear complementarity problem can easily be transformed into the solution of systems of semismooth equations $f(x) = 0$, where

$$f(x) = \begin{bmatrix} \psi(x_1, p_1(x)) \\ \dots\dots\dots\dots \\ \psi(x_n, p_n(x)) \end{bmatrix} \tag{4.24}$$

and $\psi : R^n \to R$ is a semismooth function satisfying $\psi(u_1, u_2) = 0$ if and only if $u_1 \geq 0$, $u_2 \geq 0$ and $u_1 u_2 = 0$. This property has, for example, the Pang function

$$\psi(u) = \min(u_1, u_2),$$

which is semismooth by Corollary 4.3 (since $\min(u_1, u_2) = -\max(-u_1, -u_2)$). A disadvantage of the Pang function is the fact that it does not guarantee continuos differentiability of mapping $\varphi = (1/2) f^T f$, which is used in a line search procedure. More advantageous is the Fischer-Burmeister function

$$\psi(u) = \sqrt{u_1^2 + u_2^2} - (u_1 + u_2). \tag{4.25}$$

Lemma 4.1 *Function $\psi : R^2 \to R$ defined by formula (4.25) is continuously differentiable in $R^2 \backslash \{0\}$ and semismooth at point 0, where $\partial_B \psi(0) = S(-e, 1)$, $\partial \psi(0) = \overline{B(-e, 1)}$ and $e = [1, 1]^T$ ($S(u, \varepsilon)$ is a circular line and $\overline{B(u, \varepsilon)} = \text{conv} S(u, \varepsilon)$ is a circle with the center u and radius ε). Equality $\psi(u_1, u_2) = 0$ holds if and only if $u_1 \geq 0$, $u_2 \geq 0$ and $u_1 u_2 = 0$. The square of function ψ is continuously differentiable in R^2.*

Proof. Continuous differentiability of function ψ in $R^2 \backslash \{0\}$ is obvious. For $u \in R^2 \backslash \{0\}$ one has

$$\nabla \psi(u) = \begin{bmatrix} \frac{u_1}{\sqrt{u_1^2 + u_2^2}} - 1 \\ \frac{u_2}{\sqrt{u_1^2 + u_2^2}} - 1 \end{bmatrix}. \tag{4.26}$$

Semismoothness of function ψ at point 0 follows from Theorem 4.4, since function ψ is convex (it is a sum of the Euclidean norm $\sqrt{u_1^2 + u_2^2}$ and the linear function $-(u_1 + u_2)$). Consider a sequence $\{u_i\}$, where $u_i = [t_i \cos \varphi_i, t_i \sin \varphi_i]^T$ and $t_i \downarrow 0$. Then $\nabla \psi(u_i) = [\cos \varphi_i - 1, \sin \varphi_i - 1]^T$ holds and sequence $\{\nabla \psi(u_i)\}$ has the limit $[\cos \varphi - 1, \sin \varphi - 1]^T$ if and only if $\varphi_i \to \varphi$. This implies

$$\partial_B \psi(0) = \bigcup_{\varphi \in [0, 2\pi]} [\cos \varphi - 1, \sin \varphi - 1]^T = S(-e, 1)$$

and

$$\partial\psi(0) = \text{conv } \partial_B\psi(0) = \text{conv } S(-e,1) = \overline{B(-e,1)}.$$

If $u_1 < 0$, one has

$$\psi(u) = \sqrt{|u_1|^2 + u_2^2} + |u_1| - u_2 \geq |u_2| + |u_1| - u_2 > 0$$

(the same result is obtained for $u_2 < 0$). If $u_1 > 0$, $u_2 > 0$, we can write

$$\psi(u) = \sqrt{u_1^2 + u_2^2} - (u_1 + u_2) < \sqrt{u_1^2 + 2u_1u_2 + u_2^2} - (u_1 + u_2) = 0.$$

If $u_1 = 0$ a $u_2 > 0$, one has

$$\psi(u) = |u_2| - u_2 = 0$$

(the same result is obtained for $u_1 > 0$ and $u_2 = 0$). Equality $\psi(0) = 0$ is obvious. The square of function ψ can be expressed in the form

$$\psi^2(u) = u_1^2 + u_2^2 + (u_1 + u_2)^2 - 2(u_1 + u_2)\sqrt{u_1^2 + u_2^2}.$$

This function is continuously differentiable in $R^2\backslash\{0\}$ and is continuously differentiable at point 0 if and only if the function $\overline{\psi}(u) = (u_1 + u_2)\sqrt{u_1^2 + u_2^2}$ is continuously differentiable at point 0. But

$$\lim_{|u|\to 0} \frac{\overline{\psi}(u) - \overline{\psi}(0)}{|u|} = \lim_{|u|\to 0} \frac{(u_1 + u_2)\sqrt{u_1^2 + u_2^2}}{\sqrt{u_1^2 + u_2^2}} = 0$$

so that $\overline{\psi}$ is differentiable at point 0 and $\nabla\overline{\psi}(0) = 0$ holds. Continuity of partial derivative $\partial\overline{\psi}/\partial u_1$ at point 0 follows from the inequality

$$\begin{aligned}
\left|\frac{\partial\overline{\psi}(u)}{\partial u}\right| &= \left|\frac{u_1}{\sqrt{u_1^2 + u_2^2}}(u_1 + u_2) + \sqrt{u_1^2 + u_2^2}\right| \\
&\leq \frac{|u_1|}{\sqrt{u_1^2 + u_2^2}}|u_1 + u_2| + \sqrt{u_1^2 + u_2^2} \leq |u_1 + u_2| + \sqrt{u_1^2 + u_2^2}
\end{aligned}$$

and from the fact that the right hand side of this inequality converges to zero whenever $u \to 0$ (the same results is obtained for partial derivative $\partial\overline{\psi}/\partial u_2$). $\diamond$

Theorem 4.11 *Let mapping $p : R^n \to R^n$ be continuously differentiable at point $x \in R^n$. Let $f : R^n \to R^n$ be the mapping defined by formula (4.24), where $\psi : R^2 \to R$ is the function defined by (4.25). Then:*

(a) Mapping f is semismooth at point x.

(b) One has $\partial_B f(x) \subset [\partial_B f_1(x), \ldots, \partial_B f_n(x)]^T$, where

$$\partial_B f_i(x) = \nabla f_i(x) = \left(\frac{x_i}{\sqrt{x_i^2 + p_i^2(x)}} - 1\right) e_i + \left(\frac{p_i(x)}{\sqrt{x_i^2 + p_i^2(x)}} - 1\right) \nabla p_i(x), \qquad (4.27)$$

if $x_i^2 + p_i^2(x) \neq 0$ *and*

$$\partial_B f_i(x) = \bigcup_{\varphi \in [0,2\pi]} [(\cos\varphi - 1)e_i + (\sin\varphi - 1)\nabla p_i(x)], \tag{4.28}$$

if $x_i^2 + p_i^2(x) = 0$.

(c) Function $\varphi = (1/2) f^T f$ *is continuously differentiable at point* x.

Proof. (a) Semismoothness of mapping f follows from Theorem 4.2 and Theorem 4.5, since $f_i(x) = \psi(x_i, p_i(x))$, function ψ is semismooth by Lemma 4.1 and mapping p is continuously differentiable.
(b) By Lemma 4.1, function $\psi(x_i, p_i)$ is continuously differentiable if $x_i^2 + p_i^2 \neq 0$. Formula (4.27) follows from (4.26) by the rule for the differentiation of a composite function. If $x_i^2 + p_i^2 = 0$, we can use the same limiting process as in Lemma 4.1 so that

$$\partial_B f_i(x) = [e_i, \nabla p_i(x)] \partial_B \psi(0) = [e_i, \nabla p_i(x)] S(-e, 1),$$

which gives (4.28).
(c) One has

$$\varphi(x) = \frac{1}{2} f^T(x) f(x) = \frac{1}{2} \sum_{i=1}^n \psi^2(x_i, p_i(x)).$$

Mapping p is continuously differentiable. By Lemma 4.1, the square of function ψ is continuously differentiable, so that function φ is also continuously differentiable. $\diamond$

Theorem 4.11 gives one possibility for solving nonlinear complementarity problems. A nonlinear complementarity problem is transformed to the system of nonsmooth equations (4.24), which is solved by Algorithm 4.1. Using Remark 4.8, we can choose $J_k \in \partial_b f(x_k)$, where the set $\partial_b f(x_k) = [\partial_B f_1(x_k), \dots, \partial_B f_n(x_k)]^T$ can be determined by (4.27)-(4.28). Function $\varphi = (1/2) f^T f$, used in a line search procedure, is continuously differentiable in this case.

We shall show another application of nonsmooth equations. Consider the nonlinear programming problem: Find a minimum of continuously differentiable function $f : R^n \to R$ on the set defined by constraints $c_i(x) \leq 0$, $1 \leq i \leq m$, where $c : R^n \to R^m$, is a continuously differentiable mapping. If constraint qualification is satisfied, then a solution of this problem has to satisfy the following conditions

$$\nabla f(x) + \sum_{i=1}^m \lambda_i \nabla c_i(x), \tag{4.29}$$

$$\left. \begin{array}{l} -c_i(x) \geq 0, \quad \lambda_i \geq 0, \\ \lambda_i c_i(x) = 0, \quad 1 \leq i \leq m \end{array} \right\} \tag{4.30}$$

(Theorem 1.3). Conditions (4.30) are in fact the nonlinear complementarity conditions (4.23). Therefore, we can set up the system of $n + m$ nonsmooth equations

$$F(x, \lambda) \triangleq \left[\begin{array}{l} \nabla f(x) + \sum_{i=1}^m \lambda_i \nabla c_i(x) \\ \psi(\lambda_1, -c_1(x)) \\ \dots\dots\dots\dots\dots\dots\dots\dots \\ \psi(\lambda_m, -c_m(x)) \end{array} \right] = 0, \tag{4.31}$$

where ψ is the Fischer-Burmeister function (4.25). Function $F : R^{n+m} \to R^{n+m}$ is semismooth and function $\varphi = (1/2)F^T F$ is continuously differentiable so that system of equations (4.31) can be solved by Algorithm 4.1.

5 Nonsmooth optimization

5.1 Bundle methods

We will assume that function $f : R^n \to R$ is locally Lipschitz and that we are able to compute a subgradient $g \in \partial f(x)$ at any point $x \in R^n$. Since a locally Lipschitz function is differentiable almost everywhere by the Rademacher theorem, then usually $g = \nabla f(x)$. A special feature of nonsmooth optimization methods is the fact that the gradient $\nabla f(x)$ can change discontinuously and it may not be small at a neighborhood of an extremum of function f. For this reason, values $f(x_k)$, $g(x_k) \in \partial f(x_k)$ at single point x_k do not suffice for describing local behaviour of function f, but a complete bundle of values

$$f_j = f(y_j), \quad g_j \in \partial f(y_j), \tag{5.1}$$

obtained at trial points y_j, $j \in \mathcal{J}_k \subset \{1, \ldots, k\}$, is necessary. These values serve for the construction of the piecewise linear function

$$f_L^k(x) = \max_{j \in \mathcal{J}_k}(f_j + g_j^T(x - y_j)) = \max_{j \in \mathcal{J}_k}(f_j^k + g_j^T(x - x_k)) = \max_{j \in \mathcal{J}_k}(f(x_k) + g_j^T(x - x_k) - \alpha_j^k),$$

where

$$\begin{aligned} f_j^k &= f_j + g_j^T(x_k - y_j), && (5.2)\\ \alpha_j^k &= f(x_k) - f_j^k, && (5.3) \end{aligned}$$

for $j \in \mathcal{J}_k$. This piecewise linear function is majorized by function f in the convex case.

Theorem 5.1 *Let function $f : R^n \to R$ be convex. Then $\alpha_j^k \geq 0$ $\forall j \in \mathcal{J}_k$ and $f(x) \geq f_L^k(x)$ $\forall x \in R^n$ hold for an arbitrary index k.*

Proof. Since $g_j \in \partial f(y_j)$, one has $f(x) \geq f_j + g_j^T(x - y_j)$, $\forall j \in \mathcal{J}_k$, by Theorem 2.13 (d) so that (5.2) implies $f(x_k) \geq f_j^k$, which gives $\alpha_j^k \geq 0$ by (5.3). Moreover

$$f(x) \geq \max_{j \in \mathcal{J}_k}(f_j + g_j^T(x - y_j)) = f_L^k(x).$$

◇

Theorem 5.1 does not hold in the case when function f is not convex. To guarantee applicability of piecewise linear model $f_L^k(x)$ in this case, it is necessary to choose a different definition of numbers $\alpha_j^k \geq 0$, $j \in \mathcal{J}_k$. One possibility is to set

$$\alpha_j^k = \max\left(|f(x_k) - f_j^k|, \gamma\|x_k - y_j\|^\nu\right)$$

for $j \in \mathcal{J}_k$, where $\gamma > 0$ and $\nu \geq 1$. Since it would be necessary to store points y_j, $j \in \mathcal{J}_k$, we use the fact that for $j \in \mathcal{J}_k$ one has

$$\|x_k - y_j\| \le \|x_j - y_j\| + \sum_{i=j}^{k-1} \|x_{i+1} - x_i\| \triangleq s_j^k \tag{5.4}$$

and numbers $\alpha_j^k \ge 0$, $j \in \mathcal{J}_k$ are determined by the formula

$$\alpha_j^k = \max\left(|f(x_k) - f_j^k|, \gamma(s_j^k)^\nu\right). \tag{5.5}$$

Function f_L^k is not suitable itself for determining a new approximation to the required minimum, since its minimum may not exist (f_L^k is piecewice linear) and, in case it exists, it can be too far from the minimum of function f. Therefore, a regularizing quadratic term has to be added to function f_L^k. Using this approach, we obtain the piecewise quadratic function

$$\begin{aligned} f_Q^k(x) &= \frac{1}{2}(x - x_k)^T G_k (x - x_k) + f_L^k(x) \\ &= \frac{1}{2}(x - x_k)^T G_k (x - x_k) + \max_{j \in \mathcal{J}_k}(f(x_k) + g_j^T(x - x_k) - \alpha_j^k), \end{aligned}$$

where G_k is a symmetric positive definite matrix.

Since, from a practical point of view, we can only work with limited-size bundles, where $|\mathcal{J}_k| \le m$, say ($|\mathcal{J}_k|$ is the cardinality of set $\mathcal{J}_k$), the set $\mathcal{J}_k$ is usually determined in such a way that $\mathcal{J}_k = \{1, \ldots, k\}$, whenever $k \le m$, and $\mathcal{J}_{k+1} = \mathcal{J}_k \cup \{k+1\} \backslash \{k+1-m\}$, whenever $k \ge m$.

If $\mathcal{J}_k \ne \{1, \ldots, k\}$, then one possibility guaranteeing the global convergence of the bundle method is the use of aggregate values f_a^k, g_a^k, s_a^k, which accumulate information from previous iterations. These values represent a linear function, which is added to the set of linear functions contained in the bundle. New aggregate values f_a^{k+1}, g_a^{k+1}, s_a^{k+1}, are computed by (5.23) from values $\tilde{f}_a^k$, $\tilde{g}_a^k$, $\tilde{s}_a^k$, obtained by solving quadratic programmig subproblem (5.7)-(5.9).

If we use aggregate values, the piecewise quadratic function has the form

$$f_Q^k(x) = \frac{1}{2}(x - x_k)^T G_k (x - x_k) + \max_{j \in \mathcal{J}_k}(f_L^k(x), f(x_k) + (x - x_k)^T g_a^k - \alpha_a^k),$$

where

$$\alpha_a^k = \max\left(|f(x_k) - f_a^k|, \gamma(s_a^k)^\nu\right). \tag{5.6}$$

The minimum of this function can be expressed in the form $x_{k+1} = x_k + d_k$, where the direction vector d_k is the solution of quadratic programming problem: Minimize function

$$\frac{1}{2} d^T G_k d + v \tag{5.7}$$

on the set determined by the constraints

$$-\alpha_j^k + d^T g_j \le v, \quad j \in \mathcal{J}_k, \tag{5.8}$$
$$-\alpha_a^k + d^T g_a^k \le v \tag{5.9}$$

(minimization proceeds over all pairs $(d, v) \in R^{n+1}$ satisfying inequalities (5.8), (5.9)).

Theorem 5.2 *Solution of the problem (5.7)-(5.9) can be expressed in the form*

$$
\begin{aligned}
d_k &= -G_k^{-1}\tilde{g}_a^k, && (5.10)\\
v_k &= -d_k^T G_k d_k - \tilde{\alpha}_a^k, && (5.11)
\end{aligned}
$$

where

$$
\begin{aligned}
\tilde{g}_a^k &= \sum_{j\in\mathcal{J}_k} \lambda_j^k g_j + \lambda_a^k g_a^k, && (5.12)\\
\tilde{\alpha}_a^k &= \sum_{j\in\mathcal{J}_k} \lambda_j^k \alpha_j^k + \lambda_a^k \alpha_a^k && (5.13)
\end{aligned}
$$

and where the Lagrange multipliers λ_j^k, $j \in \mathcal{J}_k$, λ_a^k, are solutions of the dual quadratic programming problem: Minimize function

$$\frac{1}{2}\left(\sum_{j\in\mathcal{J}_k} \lambda_j g_j + \lambda_a g_a^k\right)^T G_k^{-1}\left(\sum_{j\in\mathcal{J}_k} \lambda_j g_j + \lambda_a g_a^k\right) + \sum_{j\in\mathcal{J}_k} \lambda_j \alpha_j^k + \lambda_a \alpha_a^k \tag{5.14}$$

on the set determined by the constraints

$$\left.\begin{array}{l} \lambda_j \geq 0, \quad j \in \mathcal{J}_k, \quad \lambda_a \geq 0, \\ \sum_{j\in\mathcal{J}_k} \lambda_j + \lambda_a = 1. \end{array}\right\} \tag{5.15}$$

The minimum value of function (5.14), which corresponds to the solution of problem (5.14)-(5.15), is

$$w_k = \frac{1}{2}(\tilde{g}_a^k)^T G_k^{-1}\tilde{g}_a^k + \tilde{\alpha}_a^k = -v_k - \frac{1}{2}(\tilde{g}_a^k)^T G_k^{-1}\tilde{g}_a^k. \tag{5.16}$$

Proof. Function (5.7) is convex, since matrix G_k is positive definite. Constraints (5.8)-(5.9) are linear and, therefore, also convex so that pair $(d_k, v_k) \in R^{n+1}$ is, by Theorem 1.3, a solution of problem (5.7)-(5.9) if and only if there are Lagrange multipliers $\lambda_j^k \geq 0$, $j \in \mathcal{J}_k$, $\lambda_a^k \geq 0$, such that

$$\begin{bmatrix} G_k d_k \\ 1 \end{bmatrix} + \sum_{j\in\mathcal{J}_k} \lambda_j^k \begin{bmatrix} g_j \\ -1 \end{bmatrix} + \lambda_a^k \begin{bmatrix} g_a^k \\ -1 \end{bmatrix} = 0, \tag{5.17}$$

where

$$
\begin{aligned}
\lambda_j^k > 0 \quad &\Rightarrow \quad -\alpha_j^k + d_k^T g_j = v_k,\\
\lambda_a^k > 0 \quad &\Rightarrow \quad -\alpha_a^k + d_k^T g_a^k = v_k
\end{aligned}
$$

(complementarity conditions). The last equation of system (5.17) implies that

$$\sum_{j\in\mathcal{J}_k} \lambda_j^k + \lambda_a^k = 1.$$

Therefore, one has (5.10), (5.10) and (5.15). If we use notation (5.12)-(5.13) and the complementarity conditions, we can write

$$-\tilde{\alpha}_a^k + d_k^T \tilde{g}_a^k = v_k,$$

which together with (5.10) gives (5.11). It remains to prove that the Lagrange multipliers $\lambda_j^k \geq 0$, $j \in \mathcal{J}_k$, $\lambda_a^k \geq 0$ are solution of problem (5.14)-(5.15). This problem is also convex so that, by Theorem 1.3, numbers $\lambda_j^k \geq 0$, $j \in \mathcal{J}_k$, $\lambda_a^k \geq 0$, are its solution if and only if there are Langrange multipliers v_k (corresponding to equality in (5.15)) and $\mu_j^k \geq 0$, $j \in \mathcal{J}_k$, $\mu_a^k \geq 0$ (corresponding to inequalities in (5.15)) such that

$$\begin{aligned} -(g_j)^T d_k + \alpha_j^k + v_k - \mu_j^k &= 0, \quad j \in \mathcal{J}_k, \\ -(g_a^k)^T d_k + \alpha_a^k + v_k - \mu_a^k &= 0, \end{aligned}$$

where $\lambda_j^k \mu_j^k = 0$, $j \in \mathcal{J}_k$, $\lambda_a^k \mu_a^k = 0$ (we have used notations (5.10) and (5.12) to simplify these formulas). However, the last equalities are exactly inequalities (5.8), (5.9), since $\mu_j^k \geq 0$, $j \in \mathcal{J}_k$, $\mu_a^k \geq 0$, and conditions $\lambda_j^k \mu_j^k = 0$, $j \in \mathcal{J}_k$, $\lambda_a^k \mu_a^k = 0$ are equivalent to complementarity conditions for problem (5.7)-(5.9). ◇

Remark 5.1 Notice that constraint (5.9) need not be used if $\mathcal{J}_k = \{1, \ldots, k\}$, since it is a linear combination of constraints (5.8) in this case. However, then $\lambda_a^k = 0$ in (5.11)-(5.14).

Remark 5.2 Besides values (5.12), (5.13), the aggregate values

$$\tilde{f}_a^k = \sum_{j \in \mathcal{J}_k} \lambda_j^k f_j^k + \lambda_a^k f_a^k, \tag{5.18}$$

$$\tilde{s}_a^k = \sum_{j \in \mathcal{J}_k} \lambda_j^k s_j^k + \lambda_a^k s_a^k. \tag{5.19}$$

are defined using the Lagrange multipliers $\lambda_j^k \geq 0$, $j \in \mathcal{J}_k$, $\lambda_a^k \geq 0$.

Having the direction vector d_k determined, it is necessary to compute a new approximation of the minimum of function f. To guarantee the global convergence of a bundle method, it is not possible to just set $x_{k+1} = x_k + d_k$, but it is neccessary to use a more complex procedure which generate two points

$$\begin{aligned} x_{k+1} &= x_k + t_L^k d_k, \\ y_{k+1} &= x_k + t_R^k d_k, \end{aligned}$$

where $0 \leq t_L^k \leq t_R^k \leq 1$ are steplenghts. Steplengths are chosen in such a way (Algorithm 5.2), that exactly one of the two possibilities given by Definition 5.1 and Definition 5.2 occurs. In both of these definitions we use the notation

$$\beta_{k+1} = \max\left(|f(x_k) - f_{k+1} - (x_k - y_{k+1})^T g_{k+1}|, \gamma |x_k - y_{k+1}|^\nu\right) \tag{5.20}$$

and constants $0 < \sigma_L < \sigma_T < \sigma_R < 1$, $0 < \sigma_A < \sigma_R - \sigma_T$, $0 < \tau < 1$ and $D > 0$.

Definition 5.1 *(Descent step) The descent step is defined by the conditions* $t_R^k = t_L^k > 0$,

$$f(x_{k+1}) \le f(x_k) - \sigma_L t_L^k w_k \tag{5.21}$$

and either $t_L^k \ge \tau$ *or* $\beta_{k+1} > \sigma_A w_k$.

Definition 5.2 *(Zero step) The zero step is defined by the conditions* $t_R^k > t_L^k = 0$,

$$d_k^T g_{k+1} \ge \beta_{k+1} - \sigma_R w_k \tag{5.22}$$

and $\|y_{k+1} - z_{k+1}\| \le D$, *where* z_{k+1} *is any point such that* $f(z_{k+1}) \le f(x_k)$.

Having point x_{k+1} determined, it is necessary to transform bundle and aggregate values to this point. This is realized by the formulas

$$\left.\begin{array}{l}
f_j^{k+1} = f_j^k + (x_{k+1} - x_k)^T g_j, \quad j \in J_k \\
f_a^{k+1} = \tilde{f}_a^k + (x_{k+1} - x_k)^T \tilde{g}_a^k \\
f_{k+1}^{k+1} = f_{k+1} + (x_{k+1} - y_{k+1}) g_{k+1} \\
g_a^{k+1} = \tilde{g}_a^k \\
s_j^{k+1} = s_j^k + \|x_{k+1} - x_k\|, \quad j \in J_k \\
s_a^{k+1} = \tilde{s}_a^k + \|x_{k+1} - x_k\| \\
s_{k+1}^{k+1} = \|x_{k+1} - y_{k+1}\|
\end{array}\right\} \tag{5.23}$$

It remains to specify the way for determining matrices G_k. To assure the global convergence of a bundle method, we use the following assumption.

Assumption 5.1 *Matrices* G_k *are uniformly positive definite and uniformly bounded (their eigenvalues lie in the compact interval that does not contain zero). If the k-th step is a zero step, then* $h^T G_{k+1}^{-1} h \le h^T G_k^{-1} h$, $\forall h \in R^n$.

Now we are in a position to describe the basic algorithm of bundle methods.

Algorithm 5.1

Data $\varepsilon > 0$, $\gamma > 0$, $\nu \ge 1$, $m \ge 1$.

Step 1 (Initiation). Determine an initial point $x_1 \in R^n$ and an initial symmetric positive definite matrix G_1. Set $y_1 = x_1$ and compute values $f_1 = f(y_1)$, $g_1 \in \partial f(y_1)$. Set $s_1^1 = s_a^1 = 0$, $f_1^1 = f_a^1 = f_1$, $g_1^1 = g_a^1 = g_1$, $J_1 = \{1\}$ and $k = 1$.

Step 2 (Direction vector). Find the solution of quadratic programming subproblem (5.7)-(5.9) (constraint (5.9) is used only if $J_k \ne \{1, \dots, k\}$). Thus the Lagrange multipliers λ_j^k, $j \in J_k$ and λ_a^k ($\lambda_a^k \ne 0$ only if $J_k \ne \{1, \dots, k\}$), aggregate values $\tilde{g}_a^k$, $\tilde{\alpha}_a^k$, $\tilde{f}_a^k$, $\tilde{s}_a^k$, direction vector d_k and values v_k, w_k (Theorem 5.2) are obtained. If $w_k \le \varepsilon$, terminate the computation.

Step 3 (Line search). By using Algorithm 5.2, determine steplengths t_L^k, t_R^k to obtain either a descent step (Definition 5.1) or a zero step (Definition 5.2). Set $x_{k+1} = x_k + t_L d_k$, $y_{k+1} = x_k + t_R d_k$ and compute values $f_{k+1} = f(y_{k+1})$, $g_{k+1} \in \partial f(y_{k+1})$.

Step 4 (Update). Compute transformed values by (5.23) and determine a matrix G_{k+1} satisfying Assumption 5.1. If $|\mathcal{J}_k| < m$, set $\mathcal{J}_{k+1} = \mathcal{J}_k \cup \{k+1\}$. If $|\mathcal{J}_k| = m$, set $\mathcal{J}_{k+1} = \mathcal{J}_k \cup \{k+1\} \backslash \{k+1-m\}$. Set $k := k+1$ and go to Step 2.

Remark 5.3 Set $\mathcal{J}_{k+1}$ can be also determined in a different way than it is described in Step 4 of Algorithm 5.1. In principle, it should contain a sufficient number of points, but inclusion $k+1 \in \mathcal{J}_{k+1}$ is necessary.

Line search (Step 3 of Algorithm 5.1) is a relatively complicated procedure, which is introduced in the form of the following algorithm. To simplify the notation, we omit index k and replace index $k+1$ by the symbol $+$.

Algorithm 5.2

Data $0 < \sigma_L < \sigma_T < \sigma_R < 1$, $0 < \sigma_A < \sigma_R - \sigma_T$, $\gamma > 0$, $\nu \geq 1$, $0 < \kappa < 1/2$, $0 < \tau < 1/2$, $D > 0$.

Input $x \in R^n$, $d \in R^n$, $f = f(x)$, $w > 0$.

Step 1 (Initiation). Set $t^1 = 1$, $t_A^1 = 0$, $t_U^1 = 1$ and $i = 1$.

Step 2 (New values). Compute values $f^i = f(x + t^i d)$, $g^i \in \partial f(x + t^i d)$ and

$$\beta^i = \max\left(|f - f^i + t^i d^T g^i|, \gamma(t^i\|d\|)^\nu\right).$$

If $f^i \leq f - \sigma_T t^i w$, set $t_A^i = t^i$. Otherwise set $t_U^i = t^i$.

Step 3 (Descent step). If $f^i \leq f - \sigma_L t^i w$ and either $t^i \geq \tau$ or $\beta^i > \sigma_A w$, then set $t_R = t_L = t^i$, $t_A = t_A^i$, $\beta^+ = \beta^i$ and terminate the computation.

Step 4 (Zero step). If $d^T g^i \geq \beta^i - \sigma_R w$ and $(t^i - t_A^i)\|d\| \leq D$, set $t_R = t^i$, $t_L = 0$, $t_A = t_A^i$, $\beta^+ = \beta^i$ and terminate the computation.

Step 5 (Update). Choose $t^{i+1} \in [t_A^i + \kappa(t_U^i - t_A^i), t_U^i - \kappa(t_U^i - t_A^i)]$, set $i := i+1$ and go to Step 2.

Theorem 5.3 *Let function $f : R^n \to R$ be locally Lipschitz and let for an arbitrary sequence $t^i \downarrow 0$*

$$\limsup_{\substack{g^i \in \partial f(x+t^i d) \\ i \to \infty}} d^T g^i \geq \liminf_{i \to \infty} \frac{f(x + t^i d) - f(x)}{t^i} \tag{5.24}$$

holds. Then Algorithm 5.2 finds, after a finite number of steps, steplengths t_L, t_R, t_A such that exactly one of two following cases occurs for points $x^+ = x + t_L d$, $y^+ + t_R d$, $z^+ = z + t_A d$:

(a) Descent step: One has $t_R = t_L > 0$,

$$f(x^+) \leq f(x) - \sigma_L t_L w$$

and either $t_L \geq \tau$ or $\beta^+ > \sigma_A w$.

(b) Zero step: One has $t_R > t_L = 0$,

$$d^T g(y^+) \geq \beta^+ - \sigma_R w,$$

$\|y^+ - z^+\| \leq D$ and $f(z^+) \leq f(x)$.

In both cases, we use the notation

$$\beta^+ = \max\left(|f(x) - f(y^+) - (x - y^+)^T g^+|, \gamma\|x - y^+\|)^\nu\right).$$

Proof. The algorithm terminates either in Step 3, then (a) holds obviously, or in Step 4 when (b) is satisfied. It remains to prove that the algorithm terminates after a finite number of steps. For contradiction purposes, we will assume that the algorithm does not terminate after a finite number of steps. Let $\{t^i\}$, $\{t_A^i\}$, $\{t_U^i\}$, $\{g^i\}$, $\{\beta^i\}$ be the sequences of values used in the algorithm (so that either $t^i = t_A^i$ or $t^i = t_U^i$). Since $t_A^i \le t_A^{i+1} \le t_U^{i+1} \le t_U^i$ and $t_U^{i+1} - t_A^{i+1} \le (1-\kappa)(t_U^i - t_A^i)$ for all indices i, there is necessarily a value $t^* \ge 0$ such that $t_A^i \uparrow t^*$, $t_U^i \downarrow t^*$ and $t_i \to t^*$. Moreover $(t^i - t_A^i)\|d\| \le D$ holds for sufficiently large indices. Denote $S = \{t \ge 0 : f(x+td) \le f - \sigma_T t w\}$. Since $\{t_A^i\} \subset S$, $t_A^i \uparrow t^*$ and function f is continuous, inequality

$$f(x + t^* d) \le f - \sigma_T t^* w \tag{5.25}$$

necessarily holds so that $t^* \in S$. Let $I = \{i : t^i \notin S\}$. First we show that the set I is infinite. If there exists an index $\bar{i} \in I$ such that $t^i \in S$, $\forall i > \bar{i}$, then $t_U^{\bar{i}} = t_U^i \downarrow t^*$ $\forall i > \bar{i}$ or $t^* = t_U^{\bar{i}} \notin S$ would have to hold, which is in contradiction to $t^* \in S$. Therefore, set I is infinite and one has $f(x + t^i d) > f - \sigma_T t^i w$, $\forall i \in I$, which together with (5.25) gives

$$\frac{f(x + t^i d) - f(x + t^* d)}{t^i - t^*} > -\sigma_T w, \quad \forall i \in I.$$

Using assumption (5.24) we obtain

$$-\sigma_T w \le \liminf_{i \xrightarrow{I} \infty} \frac{f(x + t^* d + (t^i - t^*) d) - f(x + t^* d)}{t^i - t^*} \le \limsup_{i \xrightarrow{I} \infty} d^T g^i. \tag{5.26}$$

Now we investigate two cases.
(a) Let $t^* > 0$. By (5.25), $f(x + t^i d) \le f - \sigma_L t^i w$ holds for sufficiently large indices, since $\sigma_L < \sigma_T$, $t^i \to t^*$ and function f is continuous. Since the algorithm does not terminate, both $\beta^i \le \sigma_A w$ (Step 3 of the algorithm) and $\beta^i - d^T g^i > \sigma_R w$ (Step 4 of the algorithm) have to hold for sufficiently large indices, which together gives

$$d^T g^i < \beta^i - \sigma_R w \le -(\sigma_R - \sigma_A) w < -\sigma_T w$$

(since $w > 0$), but which is for $i \in I$ (I is infinite) in contradiction to (5.26).
(b) Let $t^* = 0$. Then $t^i \to 0$ implies $\beta^i \to 0$ (since function f is continuous and subgradients g^i are bounded in a neighborhood of point x by Theorem 3.2 (a)). Since the algorithm does not terminate, $\beta^i - d^T g^i > \sigma_R w$ (Step 4 of algorithm) has to hold for sufficiently large indices so that

$$\limsup_{i \xrightarrow{I} \infty} d^T g^i \le -\sigma_R w < -\sigma_T w,$$

which is again in contradiction to (5.26). $\diamond$

Remark 5.4 By Theorem 4.1, any weakly semismooth function satisfies condition (5.24). In this case, the expression on the right hand side of (5.24) is the directional derivative (which exists) and the expression on the left hand side is equal to limit (4.2).

Now we prove the global convergence of Algorithm 5.1. Since we will investigate properties of an infinite sequence of points generated by Algorithm 5.1, we will assume that $\varepsilon = 0$ (Step 2). Furthermore, we will use the following assumption

Assumption 5.2 *Function $f : R^n \to R$ is locally Lipschitz on the set $X + \overline{B(0,D)}$, where set $X = \{x \in R^n : f(x) \le f(x_1)\}$ is compact. Moreover, condition (5.24) is satisfied (e.g. if f is weakly semismooth).*

Remark 5.5 Since the value of function f does not increase in descent steps, one has $x_k \in X$ and since X is compact, sequence $\{x_k\}$ is bounded. Since $\|y_k - z_k\| \le D$ by Theorem 5.3, where $z_k \in X$, we can write $y_k \in X + \overline{B(0,D)}$. The set $X + \overline{B(0,D)}$ is compact so that sequence $\{y_k\}$ is bounded. Since function f is locally Lipschitz on $X + \overline{B(0,D)}$, sequence $\{g_k\}$ is bounded. By (5.27) (Lemma 5.1) sequence $\{\tilde{g}_a^k\}$ is also bounded. From (5.10) and Assumption 5.1, it follows that sequence $\{d_k\}$ is bounded.

Lemma 5.1 *There are numbers $\tilde{\lambda}_i^k \ge 0$, $1 \le i \le k$, $\tilde{\lambda}_1^k + \ldots + \tilde{\lambda}_k^k = 1$ such that values $\tilde{f}_a^k$, $\tilde{g}_a^k$, $\tilde{s}_a^k$, obtained in Step 2 of Algorithm 5.1, satisfy the relations*

$$\left(\tilde{f}_a^k, \tilde{g}_a^k, \tilde{s}_a^k\right) = \sum_{i=1}^{k} \tilde{\lambda}_i^k (f_i^k, g_i, s_i^k) \tag{5.27}$$

(parantheses in (5.27) mean that this equality holds for all numbers of a given triple).

Proof. Proof will be carried out by induction. Assume that $\tilde{f}_a^k$, $\tilde{g}_a^k$, $\tilde{s}_a^k$ satisfy the relation (5.27) (it is obviously true if $\mathcal{J}_k = \{1, \ldots, k\}$, when $\lambda_a^k = 0$, so that relations (5.12), (5.18), (5.19) imply (5.27) with $\tilde{\lambda}_i^k = \lambda_i^k$). Let $\lambda_i^{k+1} \ge 0$, $i \in \mathcal{J}_{k+1}$, be the Lagrange multipliers determined by solution of problem (5.7)-(5.9) (or (5.14)-(5.15)), where index k is replaced by $k+1$, and let $\lambda_i^{k+1} = 0$, $i \notin \mathcal{J}_{k+1}$. Let us set $\tilde{\lambda}_i^{k+1} = \lambda_i^{k+1} + \lambda_a^{k+1}\tilde{\lambda}_i^k$, $i \le k$ and $\tilde{\lambda}_{k+1}^{k+1} = \lambda_{k+1}^{k+1}$. Then, by (5.15), $\tilde{\lambda}_i^{k+1} \ge 0$ holds for $0 \le i \le k+1$ and one has

$$\sum_{i=1}^{k+1} \tilde{\lambda}_i^{k+1} = \sum_{i=1}^{k+1} \lambda_i^{k+1} + \lambda_a^{k+1} \sum_{i=1}^{k} \tilde{\lambda}_i^k = \sum_{i \in \mathcal{J}_{k+1}} \lambda_i^{k+1} + \lambda_a^{k+1} = 1.$$

Furthermore, by using (5.23), (5.12), (5.18), (5.19) we obtain

$$\begin{aligned}
(\tilde{f}_a^{k+1}, \tilde{g}_a^{k+1}, \tilde{s}_a^{k+1}) &= \sum_{j \in \mathcal{J}_{k+1}} \lambda_i^{k+1}(f_i^{k+1}, g_i, s_i^{k+1}) + \lambda_a^{k+1}\left(f_a^{k+1}, g_a^{k+1}, s_a^{k+1}\right) \\
&= \sum_{j \in \mathcal{J}_{k+1}} \lambda_i^{k+1}(f_i^{k+1}, g_i, s_i^{k+1}) \\
&\quad + \lambda_a^{k+1}\left(\tilde{f}_a^k + (x_{k+1} - x_k)^T \tilde{g}_a^k, \tilde{g}_a^k, \tilde{s}_a^k + \|x_{k+1} - x_k\|\right) \\
&= \sum_{i=1}^{k+1} \lambda_i^{k+1}(f_i^{k+1}, g_i, s_i^{k+1}) \\
&\quad + \lambda_a^{k+1} \sum_{i=1}^{k} \tilde{\lambda}_i^k \left(f_i^k + (x_{k+1} - x_k)^T g_i, g_i, s_i^k + \|x_{k+1} - x_k\|\right) \\
&= \left(\sum_{i=1}^{k+1} \lambda_i^{k+1} + \lambda_a^{k+1} \sum_{i=1}^{k} \tilde{\lambda}_i^k\right) (f_i^{k+1}, g_i, s_i^{k+1})
\end{aligned}$$

$$= \sum_{i=1}^{k+1} \tilde{\lambda}_i^{k+1} \left(f_i^{k+1}, g_i, s_i^{k+1} \right).$$

$\diamond$

Lemma 5.2 *If sequence $\{x_k\}$ generated by Algorithm 5.1 has a cluster point $x^* \in R^n$ and there is a subsequence $\{x_k\}_K \subset \{x_k\}$ such that $x_k \xrightarrow{K} x^*$ and $w_k \xrightarrow{K} 0$, then point x^* is a stationary point of function f (one has $0 \in \partial f(x^*)$).*

Proof. By Lemma 5.1, (5.27) holds. By Theorem 2.3, there are at most $n+2$ pairs $(g^{k,i}, s^{k,i})$, $g^{k,i} \in \partial f(y^{k,i})$, $(y^{k,i}, g^{k,i}, s^{k,i}) \in \{(y_i, g_i, s_i) : i = 1, \dots, k\}$, so that one has

$$(\tilde{g}_a^k, \tilde{s}_a^k) = \sum_{i=1}^{n+2} \lambda^{k,i} \left(g^{k,i}, s^{k,i} \right), \tag{5.28}$$

where $\lambda^{k,i} \ge 0$, $1 \le i \le n+2$, $\lambda^{k,1} + \ldots + \lambda^{k,n+2} = 1$. By Remark 5.5, vectors $y^{k,i}$, $g^{k,i}$, $1 \le i \le n+2$, are bounded so that there is a subset $\overline{K} \subset K$ such that $y^{k,i} \xrightarrow{\overline{K}} y_i^*$, $g^{k,i} \xrightarrow{\overline{K}} g_i^*$, $\lambda^{k,i} \xrightarrow{\overline{K}} \lambda_i^*$, $1 \le i \le n+2$. By Theorem 3.2 (c), one has $g_i^* \in \partial f(y_i^*)$, $1 \le i \le n+2$. Then (5.28) implies $(\tilde{g}_a^k, \tilde{s}_a^k) \to (\tilde{g}_a^*, \tilde{s}_a^*)$, where

$$(\tilde{g}_a^*, \tilde{s}_a^*) = \sum_{i=1}^{n+2} \lambda_i^* (g_i^*, s_i^*) \tag{5.29}$$

and $\lambda_i^* \ge 0$, $1 \le i \le n+2$, $\lambda_1^* + \ldots + \lambda_{n+2}^* = 1$. Moreover (5.4) implies $s^{k,i} \ge \|x_k - y^{k,i}\|$, which together with $x_k \xrightarrow{\overline{K}} x^*$, $y^{k,i} \xrightarrow{\overline{K}} y_i^*$ and $s^{k,i} \xrightarrow{\overline{K}} s_i^*$ gives

$$s_i^* \ge \|x^* - y_i^*\| \tag{5.30}$$

for $1 \le i \le n+2$. Since $w_k \xrightarrow{\overline{K}} 0$, matrices G_k are uniformly positive definite and $\tilde{\alpha}_a^k \ge 0$, it follows that $\tilde{g}_a^k \xrightarrow{\overline{K}} 0$, $\tilde{\alpha}_a^k \xrightarrow{\overline{K}} 0$ by (5.16). By (5.5), (5.6) a (5.13) we get

$$\begin{aligned}
\tilde{\alpha}_a^k &= \sum_{j \in \mathcal{J}_k} \lambda_j^k \max\left(|f(x_k) - f_j^k|, \gamma (s_j^k)^\nu \right) + \lambda_a^k \max\left(|f(x_k) - f_a^k|, \gamma (s_a^k)^\nu \right) \\
&\ge \max\left(\sum_{j \in \mathcal{J}_k} \lambda_j^k |f(x_k) - f_j^k| + \lambda_a^k |f(x_k) - f_a^k|, \gamma \left(\sum_{j \in \mathcal{J}_k} \lambda_j^k s_j^k + \lambda_a^k s_a^k \right)^\nu \right) \\
&\ge \max\left(|f(x_k) - \tilde{f}_a^k| \right), \gamma \left((\tilde{s}_a^k)^\nu \right),
\end{aligned} \tag{5.31}$$

since functions $\max(\cdot, \cdot)$ and $|\cdot|^\nu$, $\nu \ge 1$, are convex. Therefore, one has $\tilde{g}_a^k \xrightarrow{\overline{K}} 0$, $\tilde{s}_a^k \xrightarrow{\overline{K}} 0$, which by using (5.29) and (5.30) gives

$$(0, 0) = \sum_{i=1}^{n+2} \lambda_i^* (g_i^*, s_i^*)$$

and $y_i^* = x^*$, $1 \le i \le n+2$. Therefore, $g_i^* \in \partial f(y_i^*) = \partial f(x^*)$ and $0 = \lambda_1^* g_1^* + \ldots + \lambda_{n+2}^* g_{n+2}^* \in \partial f(x^*)$. $\diamond$

Remark 5.6 *If Algorithm 5.1 terminates with $\varepsilon = 0$, i.e., if $w_k = 0$ in some iteration, then point x_k has the same properties as point x^* in Lemma 5.2. One has $\tilde{g}_a^k = 0$ and $\tilde{s}_a^k = 0$, which gives $0 \in \partial f(x_k)$ as in Lemma 5.2.*

Lemma 5.3 *Let the number of descent steps in Algorithm 5.1 be finite a let the l-th iteration be the last descent step. Then point x_{l+1} is a stationary point of function f ($0 \in \partial f(x_{l+1})$ holds).*

Proof. Notice first that for $k > l$ one has $x_{k+1} = x_k$ so that (5.6) and (5.23) imply

$$\alpha_a^{k+1} = \max\left(|f(x_k) - f_a^{k+1}|, \gamma(s_a^{k+1})^\nu\right) = \max\left(|f(x_k) - \tilde{f}_a^k|, \gamma(\tilde{s}_a^k)^\nu\right),$$

which together with (5.31) gives $\alpha_a^{k+1} \le \tilde{\alpha}_a^k$. Let $0 \le \lambda \le 1$. Denote

$$\begin{aligned}
g_{k+1}(\lambda) &= \lambda g_{k+1} + (1-\lambda) g_a^{k+1} = \lambda g_{k+1} + (1-\lambda)\tilde{g}_a^k \triangleq \tilde{g}_k(\lambda),\\
\alpha_{k+1}(\lambda) &= \lambda \alpha_{k+1}^{k+1} + (1-\lambda)\alpha_a^{k+1} \le \lambda \alpha_{k+1}^{k+1} + (1-\lambda)\tilde{\alpha}_a^k \triangleq \tilde{\alpha}_k(\lambda).
\end{aligned}$$

Since w_{k+1} is a minimum of function (5.14) (with index $k+1$ replaced by k) by Theorem 5.2, we can write

$$w_{k+1} \le \frac{1}{2} g_{k+1}^T(\lambda) G_{k+1}^{-1} g_{k+1}(\lambda) + \alpha_{k+1}(\lambda) \le \frac{1}{2}\tilde{g}_k^T(\lambda) G_k^{-1}\tilde{g}_k(\lambda) + \tilde{\alpha}_k(\lambda) \triangleq w_k(\lambda)$$

for $k > l$, since $h^T G_{k+1}^{-1} h \le h^T G_k^{-1} h \ \forall h \in R^n$ for $k > l$ (Assumption 5.1). Furthermore, notice that (5.10) and (5.22) imply

$$\alpha_{k+1}^{k+1} + g_{k+1}^T G_k^{-1} \tilde{g}_a^k \le \sigma_R w_k$$

for $k > l$, since in zero steps $\alpha_{k+1}^{k+1} = \beta_{k+1}$ holds by (5.20). By successive arrangements, we obtain

$$\begin{aligned}
w_k(\lambda) &= \frac{1}{2}\tilde{g}_k^T(\lambda) G_k^{-1}\tilde{g}_k(\lambda) + \tilde{\alpha}_k(\lambda)\\
&= \frac{1}{2}(\tilde{g}_a^k)^T G_k^{-1}\tilde{g}_a^k + \tilde{\alpha}_a^k + \lambda\left(g_{k+1}^T G_k^{-1}\tilde{g}_a^k - (\tilde{g}_a^k)^T G_k^{-1}\tilde{g}_a^k + \alpha_{k+1}^{k+1} - \tilde{\alpha}_a^k\right)\\
&\quad + \lambda^2\left(g_{k+1} - \tilde{g}_a^k\right)^T G_k^{-1}(g_{k+1} - \tilde{g}_a^k)\\
&\le w_k + \lambda\sigma_R w_k - \lambda w_k + \lambda^2\left(g_{k+1} - \tilde{g}_a^k\right)^T G_k^{-1}(g_{k+1} - \tilde{g}_a^k)\\
&\le w_k + \lambda(\sigma_R w_k - w_k) + \lambda^2 M,
\end{aligned}$$

where the existence of a constant M follows from the boundedness of values g_{k+1}, $\tilde{g}_a^k$ (Remark 5.5) and from the uniform positive definiteness of matrices G_k (Assumption 5.1). The expression on the right hand side acquires the minimum for $\lambda = (1-\sigma_R)w_k/(2M)$ and its minimum value is equal to $w_k - (1-\sigma_R)^2 w_k^2/(4M)$. Therefore, one has

$$w_{k+1} \le w_k - \frac{(1-\sigma_R)^2 w_k^2}{4M}. \tag{5.32}$$

Now we can easily finish the proof. We show that $w_k \to 0$ for $k > l$. If it were not true, constant $\delta > 0$ would have to exist such that $w_k \ge \delta$, $\forall k > l$ (since sequence $\{w_k\}$ is nonincreasing for $k > l$). Then we would obtain $w_{k+1} \le w_k - (1-\sigma_R)^2\delta^2/(4M)\ \forall k > l$ from (5.32) so that $w_k < \delta$ would hold for sufficiently large indices, which is a contradiction. Since $x_k = x_{l+1}$, $\forall k > l$, one has $x_k \to x_{l+1}$, which together with $w_k \to 0$ gives $0 \in \partial f(x_{l+1})$ by Lemma 5.2. $\diamond$

Theorem 5.4 *Let function $f : R^n \to R$ satisfy Assumption 5.2. Then every cluster point of sequence $\{x_k\}$, generated by Algorithm 5.1, is a stationary point of function f.*

Proof. If the number of descent steps in Algorithm 5.1 is finite, then there is exactly one cluster point of sequence $\{x_k\}$, by Lemma 5.3, which is a stationary point of function f. Let us suppose that $x_k \stackrel{K}{\to} x^*$ (set K and point x^* exists, since sequence $\{x_k\}$ is bounded). Define the infinite set

$$\overline{K} = \left\{ k = k(i) : k(i) \geq i, i \in K, x_i = \ldots = x_{k(i)} \neq x_{k(i)+1} \right\}$$

so that the step with index $k \in \overline{K}$ is descent and $x_k \stackrel{\overline{K}}{\to} x^*$. Since sequence $\{f(x_k)\}$ is nonincreasing and bounded from below (as f is locally Lipschitz on a compact set), it has to have a limit and, therefore, $f(x_k) - f(x_{k+1}) \stackrel{\overline{K}}{\to} 0$. Since (5.21) holds for $k \in \overline{K}$, we can write

$$0 \leq \sigma_L t_L^k w_k \leq f(x_k) - f(x_{k+1})$$

so that $t_L^k w_k \stackrel{\overline{K}}{\to} 0$. By Theorem 5.3, one has $\overline{K} = K_1 \cup K_2$, where $K_1 = \{k \in \overline{K} : t_L^k \geq \tau\}$ and $K_2 = \{k \in \overline{K} : \beta_{k+1} > \sigma_A w_k\}$. If set K_1 is infinite, then $t_L^k w_k \stackrel{K_1}{\to} 0$ implies $w_k \stackrel{K_1}{\to} 0$ so that point x^* is a stationary point of function f by Lemma 5.2. If set K_1 is finite, then set K_2 has to be infinite. Assume that there is a number δ such that the set $K_3 = \{k \in K_2, w_k > \delta\}$ is infinite. Then $t_L^k w_k \stackrel{K_3}{\to} 0$ implies $t_L^k \stackrel{K_3}{\to} 0$. Assumption 5.1 and boundedness of direction vectors (Remark 5.5) imply existence of a number $\overline{M} > 0$ such that

$$\|x_{k+1} - x_k\| = t_L^k \|d_k\| =\leq t_L^k \overline{M}$$

so that $t_L^k \stackrel{K_3}{\to} 0$ implies $\|x_{k+1} - x_k\| \stackrel{K_3}{\to} 0$. Since $y_{k+1} = x_{k+1}$ holds in descent steps, we obtain $\|y_{k+1} - x_k\| \stackrel{K_3}{\to} 0$. This gives $\beta_{k+1} \stackrel{K_3}{\to} 0$, after a substitution to (5.20), by using the continuity of function f. Since $K_3 \subset K_2$, one has $0 \leq \sigma_A w_k < \beta_{k+1}$ so that $w_k \stackrel{K_3}{\to} 0$, which is in contradiction to the definition of the set K_3. Therefore, one has $w_k \stackrel{K_2}{\to} 0$ and point x^* is a stationary point of function f by Lemma 5.2. $\diamond$

Algorithm 5.1 represents a class of globally convergent bundle methods for minimization of nonsmooth function. Individual methods differ by the choice of matrix G_k. The simplest proximal bundle method uses the matrix

$$G_k = u_k I,$$

where $u_k > 0$ are weight coefficients. These weight coefficients are adaptively adjusted by certain (usually heuristic) rules so that $u_{\min} \leq u_k \leq u_{\max}$ and $u_{k+1} \geq u_k$ hold in zero steps (then Assumption 5.1 is satisfied). Matrix G_k can be also determined by quasi-Newton updates [3]. In this case, the rank one update, which satisfies Assumption 5.1, has to be used in zero steps. An advantage of variable metric bundle methods is the fact that matrices G_k contain relatively good information about a minimized nonsmooth function so that we can use small bundles (e.g. with $m = 1$ or $m = 2$), which leads to considerable saving of computational time for solving the quadratic programing problem (5.7)-(5.9).

References

[1] D.P.Bertsekas: Nonlinear Programming. Second Edition. Athena Scientific, Belmont, 1999.

[2] F.H.Clarke: Optimization and Nonsmooth Analysis. John Wiley & Sons, New York, 1983.

[3] Fletcher R.: Practical Methods of Optimization. Wiley, New York, 1987.

[4] J.B.Hiriart-Urruty, C.Lemarechal: Convex Analysis and Minimization Algorithms. Springer Verlag, Berlin, 1993.

[5] K.C.Kiwiel: Methods of Descent for Nondifferentiable Optimization. Lecture Notes in Mathematics, Vol. 1133, Springer Verlag, Berlin, 1985.

[6] M.M.Mäkelä, P. Neittaanmäki: Nonsmooth Optimization. World Scientific Publishing Co., London, 1992.

[7] J.Nocedal, S.J.Wright: Numerical Optimization. Springer Verlag, Berlin, 1999.

[8] J.Outrata, M.Kočvara, J.Zowe: Nonsmooth Approach to Optimization Problems with Equilibrium Constraints, Kluwer Academic Publishers, London, 1998.